Data-driven donation strategies: understanding and predicting blood donor deferral

Vinkenoog, M.

Data-Driven Donation Strategies

Understanding and Predicting Blood Donor Deferral

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op donderdag 15 februari 2024
klokke 11:15 uur

door

Marieke Vinkenoog
geboren te Amsterdam
in 1993

Promotor:

Dr. M. van Leeuwen	

Co-promotores:

Dr. M.P. Janssen	Sanquin Research & Universiteit Leiden
Dr. K. van den Hurk	Sanquin Research & Amsterdam UMC

Promotiecommissie:

Prof.dr. A. Plaat	
Dr. M. Arvas	Finnish Red Cross Blood Service
Prof.dr. R. Groenwold	LUMC & Universiteit Leiden
Prof.dr. E. Steyerberg	LUMC, Erasmus MC & Universiteit Leiden
Prof.dr. D. Swinkels	Radboud UMC

This PhD project was a collaboration between two research groups:

1. Department of Transfusion Technology Assessment, Sanquin Research, Amsterdam, The Netherlands
2. Explanatory Data Analysis, Leiden Institute of Advanced Computer Science, Leiden University, Leiden, The Netherlands

Research presented in this thesis was funded by a peer reviewed grant from Sanquin's Fund for Product and Process Development Cellular Products (PPOC 18-14/L2337).

Design and illustration by Lotte & Marieke Vinkenoog
Printed by Ridderprint — www.ridderprint.nl

Contents

CHAPTER

1

Introduction

Introduction

A safe and steady blood supply is essential for a healthy society, and requires a blood bank to collect and manage blood products. Blood that is collected from donors can be used for transfusion or for manufacturing various medicines – in either case, blood saves lives. Patient safety must be ensured; therefore donated blood is extensively tested for infectious diseases, and donors are not allowed to donate if the pre-donation check indicates a potential presence of blood-transmissible disease. In the Netherlands, adverse events for patients after a blood transfusion are extremely rare. [1] In addition to patient safety, donor safety is a priority for blood banks. Both from an ethical and practical perspective, it is important that people only donate blood when it does not harm them. The pre-donation screening visit therefore consists of a questionnaire and blood tests that assess both patient and donor safety.

The biggest health risk in terms of prevalence for blood donors is iron deficiency. [2, 3] With every whole blood donation, donors lose about 250 mg of iron, which is 8 to 13% of their iron stores for men and non-menstruating women, and up to 81% for menstruating women. [4] In general, iron deficiency symptoms may begin mild and vague, including fatigue, increased irritability, and difficulties with concentration. As the deficiency progresses, it can evolve into iron deficiency anemia, meaning that there is a shortage of healthy red blood cells. Symptoms will intensify as the need for iron increases. [5] A recent systematic review investigated associations between iron deficiency in whole blood donors and several health consequences related to iron deficiency. [4] Although most included studies reported a high prevalence of iron deficiency among blood donors, no clear overall association for most iron deficiency-related symptoms was found, and only restless leg syndrome and pica (the act of eating non-food items) were associated with iron deficiency in blood donors. [4]

To prevent anemia in blood donors, most blood banks implement pre-donation checking of hemoglobin or ferritin levels, proteins that transport oxygen and carbon dioxide and store iron, respectively. Donors that do not meet the eligibility criteria for safe blood donation are deferred, i.e., sent home without donating blood. Although it protects donor health, deferral is demotivating for donors, and is often a reason for donors to stop donating altogether. However, a stable donor pool is needed to maintain a steady blood supply, and because retaining existing donors is less costly than recruiting new donors, it is in blood banks' interests to keep donors healthy and motivated.

The Dutch blood bank Sanquin has data on millions of donations, including data

relevant to donor health, and measurements of hemoglobin and ferritin levels in particular. Using statistical and machine learning methods, the information contained in this data can be used to develop algorithms that may help to improve blood bank policies. Combining expert biomedical knowledge gained over decades of blood banking with new insights obtained with data science will allow blood banks to move towards data-driven donation strategies. Similar to precision medicine, blood donation could become more data-driven as well, for instance with tailored donation intervals or eligibility thresholds.

1.1 Research questions and contributions

Sanquin's mission is rooted in its commitment to being a knowledge-driven organisation that supplies life-saving products while upholding careful, responsible, and efficient processing of the voluntary contributions made by donors. Over time, Sanquin has collected large amounts of data on donors and their donations. These data can be leveraged by proper analyses, which would allow Sanquin to advance towards more data-driven donation strategies. Accurately predicting various donor outcomes holds the potential to optimise the blood bank process, for example, by anticipating donor deferral and subsequently adapting donor invitation strategies to minimise such deferrals.

In this thesis, we explore the application of data science in enhancing donor management. By employing several statistical and data science analysis techniques on blood donation data, we address research questions that bear significance for donor health monitoring and protection. The studies focus on investigating a series of research questions that have been categorised into three primary areas. What follows is a list of research questions that are studied in this thesis and their contributions to current knowledge.

Research questions on hemoglobin and ferritin levels and recovery after donation:

Q1 **Does a ferritin-based donor deferral policy prevent donors from returning with iron deficiency?**

In 2017, Sanquin implemented a new policy wherein donors' ferritin levels are measured routinely and donors with low ferritin levels are deferred for six or twelve months. We analysed changes in ferritin levels of deferred donors.

Q2 **What are determinants of variations in ferritin levels?**

Many factors that affect iron stores, including ferritin levels, are known and have been extensively researched. Most studies investigate associations from a single perspective, e.g., focussing only on donation-related variables, or only on environmental variables. In Chapter 4, we present a statistical model that integrates multiple sets of variables to give a more comprehensive overview of determinants of ferritin. Differences between donors and non-donors are investigated.

Q3 **Can we find groups of donors whose hemoglobin levels change in a similar manner over the course of their donor career?**

Some donors exhibit very stable hemoglobin levels during their donor career, while others show a declining trend. In Chapter 5, we regard these hemoglobin trajectories as time series and use clustering methods to find groups of donors with similar trends. Clustering is complicated when time series are very irregular and sparse, as in this case, with only a few data points per donor per year and no information about hemoglobin levels in between data points. Two methods to tackle the irregularities in these time series are compared.

One research question does not concern hemoglobin or ferritin at all, but rather focuses on SARS-CoV-2 antibodies. Before (and after) the pandemic, the main reasons for donor deferral were low hemoglobin or low ferritin levels. During the pandemic, however, the most common reason donors could not donate was the presence of a COVID-19 infection. The following research question is studied:

Q4 **How are individual characteristics and symptoms associated with IgG antibody response in COVID-19 recovered donors?**

During the COVID-19 pandemic, Sanquin monitored antibodies in regular donations, but also specifically repeatedly measured antibodies in donors who had undergone a COVID-19 infection. Chapter 6 presents a linear mixed-effects model relating antibody decay to characteristics such as sex, BMI, and age, as well as the presence of various COVID-19 symptoms. At the time of publication, this was the largest study describing these associations, and one of the few studying antibodies in a non-hospitalised cohort.

Research questions on hemoglobin deferral prediction:

Q5 **Can we accurately and reliably predict hemoglobin deferral based on historical data?**

Chapter 7 presents the main hemoglobin deferral prediction model that was developed, using support vector machines. We assessed the consequences for the blood supply if these models were used to guide donor invitations, as well as prediction performance. Additionally, it focuses on explaining why the model makes certain predictions, opening the 'black box' and analysing if associations learned by the model are consistent with the physiology behind hemoglobin metabolism. This research question is also relevant in Chapters 8 and 9.

Q6 **How do country-specific blood bank policies and donor demographics affect hemoglobin deferral prediction models?**

Although blood banks from many countries are working on hemoglobin deferral prediction, exchange and comparison of results is rarely done. Chapter 8 is the first publication from the international research group SanguinStats, a collaborative effort across five countries: Australia, Belgium, Finland, the Netherlands, and South Africa. In this chapter, researchers from all countries fit the same five prediction models to their blood bank data. Both prediction performance and variable importance are analysed for differences and similarities.

Q7 **Do ferritin measurements or genetic information add value to hemoglobin deferral prediction models?**

Many blood banks collect additional information that may improve predictions: specifically, the Finnish blood bank has collected information on iron-related genetic markers and in the Netherlands, ferritin measurement data are available. In Chapter 9, the added value of these predictor variables is investigated and compared.

1.2 Outline of this thesis

Following this introduction, this thesis continues with Chapter 2 – *Preliminaries*. In this chapter, all necessary background information needed to understand the work presented in this thesis is provided: both from a blood donation and a data science perspective.

Chapters 3 through 9 contain the research papers as published, starting with three papers on iron marker levels and their recovery after donations. Right in the middle, Chapter 6 interrupts the studies on hemoglobin and ferritin levels with a research paper on SARS-CoV-2 antibodies, just as the COVID-19 pandemic interrupted me in the middle of my PhD research. Chapters 7 through 9 focus on hemoglobin deferral prediction models.

The thesis is wrapped up with Chapter 10 – *Conclusions, general discussion and anticipated future research*, which summarises the results from Chapters 3 through 9, discusses overarching challenges, and proposes potential directions for future research and policies.

1.3 List of publications

The chapters in this thesis are based on the following publications. Publications are edited only for style cohesion; all content remains unchanged from the published versions.

Chapter	Publication
3	Vinkenoog M, van den Hurk K, van Kraaij M, van Leeuwen M, & Janssen MP (2020). *First results of a ferritin-based blood donor deferral policy in the Netherlands.* Transfusion 60(8), 1785-1792.
4	Vinkenoog M, de Groot R, Lakerveld J, Janssen MP, & van den Hurk K (2023). *Individual and environmental determinants of serum ferritin levels: A structural equation model.* Transfusion Medicine, 33(2), 113-122.
5	Vinkenoog M, Janssen MP, & van Leeuwen M (2019). *Challenges and limitations in clustering blood donor hemoglobin trajectories.* International workshop on advanced analysis and learning on temporal data, 72-84.
6	Vinkenoog M, Steenhuis M, Brinke AT, van Hasselt JG, Janssen MP, van Leeuwen M, Swaneveld FH, Vrielink H, van de Watering L, Quee F, van den Hurk K, Rispens T, Hogema B & van der Schoot CE (2022). *Associations between symptoms, donor characteristics and IgG antibody response in 2082 COVID-19 convalescent plasma donors.* Frontiers in immunology, 13.
7	Vinkenoog M, van Leeuwen M, & Janssen, MP (2022). *Explainable hemoglobin deferral predictions using machine learning models: Interpretation and consequences for the blood supply.* Vox Sanguinis, 117(11), 1262-1270.
8	Vinkenoog M, Toivonen J, Brits T, de Clippel D, Compernolle V, Karki S, Welvaert M, Meulenbeld A, van den Hurk K, van Rosmalen J, Lesaffre E, Arvas M & Janssen MP (2023). *An international comparison of hemoglobin deferral prediction models for blood banking.* Vox Sanguinis, 118(6), 430-439.
9	Vinkenoog M, Toivonen J, van Leeuwen M, Janssen MP & Arvas M (2023). *The added value of ferritin levels and genetic markers for the prediction of hemoglobin deferral.* Vox Sanguinis, 118(10), 825-834.

CHAPTER

2

Preliminaries

Preliminaries

With one foot in the world of blood transfusion and the other firmly planted in data science and statistics, the research in this thesis is multidisciplinary and requires an understanding of both fields. In this chapter, background knowledge on blood donation and iron metabolism, as well as explanations of some statistical and data science concepts are presented.

2.1 Blood donation in the Netherlands

Since 1998, when Dutch blood banks were merged into one organisation, Sanquin is the only organisation in the Netherlands authorised to collect human blood for transfusion and manage the blood supply. With 50 fixed collection sites and several mobile collection sites, the country is geographically well-covered.

The two main types of blood donation at Sanquin are whole blood and plasma donation. During a whole blood donation, 500 mL of blood is collected from a donor. The blood is collected as-is, hence the term whole blood. During a plasma donation or plasmapheresis procedure, blood is collected from a donor, but everything except the plasma is returned to the donor after a separation process. In 2022, Sanquin collected 313 386 plasma donations and 430 515 whole blood donations.

Iron is present in red blood cells as a constituent of hemoglobin, and so iron loss is much greater during a whole blood donation than a plasma donation. Our research on iron deficiency in donors is therefore focused on whole blood donors. Donors sign up to become either a whole blood donor or a plasma donor, and while sometimes whole blood donors become plasma donors, donors generally do not alternate between the two types of donations. There is a pool of about 300 000 whole blood donors that is continuously changing as people stop donating and new donors are recruited. Women are allowed to donate whole blood three times a year and men five times a year, provided that all eligibility criteria are met. Whole blood donations are collected by invitation only: donors receive an invitation by postal mail, e-mail, or SMS, with which they can visit the blood bank for a donation within two weeks. More recently, donors can plan their donation by making an online appointment.

2.1.1 Iron metabolism

Blood consists of four main components: plasma, erythrocytes (red blood cells), leukocytes (white blood cells), and thrombocytes (platelets). Erythrocytes make up nearly half of the blood's volume and are responsible for delivering oxygen to the whole body. About one third of the volume of an erythrocyte is filled with hemoglobin, a protein that contains iron, which can bind oxygen in the lungs to release it where needed. [6] Anemia is the condition where hemoglobin or erythrocyte levels are low, which can lead to symptoms as described earlier. [7] When the cause of the anemia is a low iron level, this is called iron deficiency anemia, and this will usually be reflected by a low hemoglobin level. [5]

However, iron is also used in other processes than hemoglobin synthesis, and sufficient hemoglobin levels do not guarantee the absence of iron deficiency. If such a state, called non-anemic iron deficiency, is left untreated, it may evolve into iron deficiency anemia, particularly after blood loss. Aside from oxygen transport, iron is required for many cellular functions, such as the replication and repair of DNA, as well as the synthesis of several enzymes and hormones. [8] Almost all cells use iron as a co-factor for biochemical activities, which is why iron deficiency can cause such a wide range of symptoms.

Approximately 70% of all iron in adults is present in hemoglobin, and 25% is stored in ferritin, a large iron storage protein found mostly in the liver. [8] When hemoglobin levels decrease, iron is released from ferritin to be used in erythropoiesis (the production of red blood cells). A small amount of ferritin is also present in blood. Under steady-state conditions, the concentration of ferritin in the blood is correlated with the size of the total iron stores of the body. However, ferritin levels are often elevated under inflammatory conditions and are therefore not a good indicator of iron stores in those circumstances. [9] This can already occur with low-grade inflammation due to obesity or high levels of air pollution. Under normal (non-inflamed) conditions, ferritin is considered a good indicator of the total body iron store. [10]

2.1.2 Iron monitoring strategies at Sanquin

Sanquin has several eligibility criteria regarding hemoglobin and ferritin levels in donors. Before every donation, a drop of capillary blood is collected from the donor's finger and used to test the concentration of hemoglobin. The European eligibility threshold for hemoglobin levels is 7.8 mmol/L for women and 8.4 mmol/L for men; slightly higher than the WHO cut-offs for anemia (120 g/L for women and 130 g/L for

men, which corresponds to 7.5 and 8.0 mmol/L, respectively). [11] The hemoglobin testing policy at Sanquin is the following: hemoglobin is measured in one drop of blood using a HemoCue device. If the hemoglobin level is below the donation threshold, the same test is repeated once or twice, and the value that gets recorded is the highest value. The policy states that if the difference between the measurements is larger than 0.3 mmol/L, a physician should be consulted. Donors whose hemoglobin level does not meet the threshold are deferred and sent home to be invited again three months later.

Pre-donation hemoglobin testing has always been standard practice at Sanquin, but as described in the previous section, hemoglobin level monitoring does not enable the detection of iron deficiency. Figure 2.1 clearly illustrates how hemoglobin and ferritin levels change after donation in 25 male donors. [12] The concentration of hemoglobin decreases for three days as the blood volume is replenished. [13] During this time, ferritin remains stable at the pre-donation level. The iron that is necessary for the synthesis of hemoglobin for the new erythrocytes is released from ferritin; hence, at three days post-donation we see hemoglobin levels starting to increase while ferritin levels start decreasing. After 56 days, the minimum donation interval has passed, and the donor should in theory be eligible for their next donation. At that point, hemoglobin levels are almost back to the pre-donation level. However, ferritin levels are still low, at only 55% of the pre-donation level. [12] If donors were to donate again at this moment, the same pattern would repeat and there would be an increased risk of iron deficiency (first without, then with anemia). More time between subsequent donations is needed for most of these donors to completely recover their iron stores.

To avoid collecting blood from donors with iron deficiency, Sanquin implemented ferritin-guided donation intervals starting October 2017. [14] A stepped-wedge approach was used to introduce the policy in all blood banks, and since November 2019 all blood collection sites apply the ferritin testing policy. Ferritin levels are now tested in all new donors as part of the new donor intake, which assesses general eligibility for someone to become a blood donor. Hemoglobin is also measured during this visit, so baseline values of both iron markers are available for all donors. Only test tubes of blood are collected during this intake, and no donation takes place. If eligible, the donor is invited for a first donation about three weeks after the intake. Apart from the new donor intake, ferritin is measured in repeat donors every fifth donation. According to WHO guidelines, ferritin levels below 15 µg/L indicate iron deficiency. [11] Donors with a ferritin level below 15 µg/L are therefore deferred from donating for one year. If the ferritin level is between 15 and 30 µg/L, donors are considered ‘at

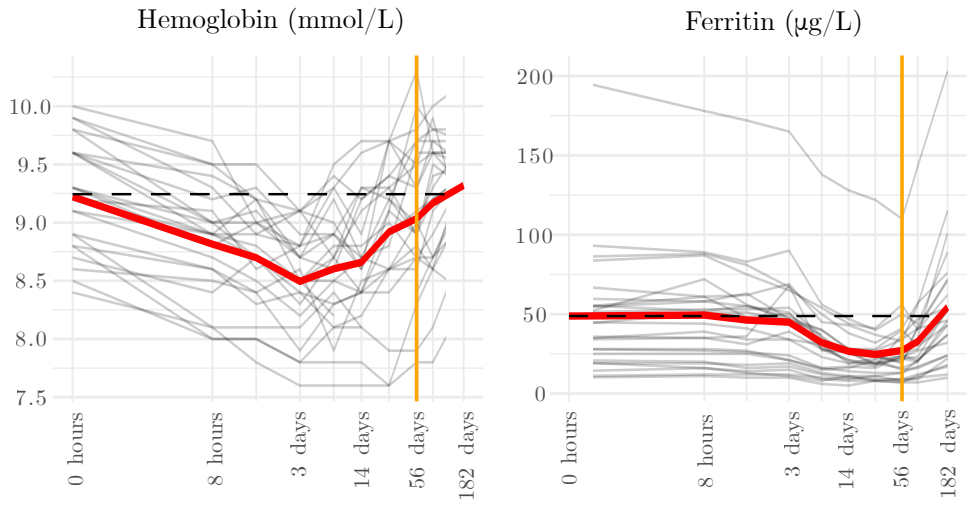

Figure 2.1: Trajectories of hemoglobin and ferritin levels after blood donation for 25 repeat male donors. Solid grey lines are individual trajectories; the red solid line is the average trajectory; the horizontal dashed line indicates the average pre-donation value, and the yellow vertical line shows the minimum donation interval for male donors. Reproduced with permission from Schotten et al., 2016. [12]

risk' of iron deficiency. To prevent the donors from returning to donate with a ferritin level below 15 µg/L, the donor is deferred for six months.

There is an important difference between measuring hemoglobin versus ferritin levels. Hemoglobin testing is done using a point-of-care device and the result is available immediately, so it can be used on-site to decide on the donor's eligibility for donation. However, measuring ferritin levels with such a device is expensive and unreliable, therefore they are measured in the donated blood after the donation has taken place. This means that when a ferritin level is measured, the actual ferritin level of the donor will reduce subsequently as a result of the donation.

After a deferral period due to low ferritin, ferritin is measured again at the next donation. If ferritin is still low after a deferral, the donation frequency is decreased (e.g., from three to two donations a year for women) or the donor can switch donation types (plasma donation might be a better choice) or decide to stop donating altogether.

2.2 Machine learning

Several chapters in this thesis describe models that predict hemoglobin deferral of blood donors. This is a form of binary classification using supervised machine learning algorithms. In those chapters, some statistical concepts are mentioned that are explained in more detail in this section.

2.2.1 Models, explanations and predictions

Machine learning methods can generally be classified as either supervised or unsupervised. In supervised learning, a model learns from examples (training data) where the true outcome is known to the model. After learning from the training data, the resulting model can be used to predict the outcome on unseen test data, for which the model does not know the true outcome. We use supervised learning methods to predict hemoglobin deferral (Chapters 7 through 9), as well as to describe COVID-19 antibody levels (Chapter 6).

In unsupervised learning, there is no outcome to predict. Instead, these models are used to find, for example, groups of either observations (clustering tasks) or variables (dimensionality reduction). We apply unsupervised learning methods to cluster hemoglobin trajectories (Chapter 5), and to group predictor variables into latent constructs for the structural equation model (Chapter 4).

An important distinction in statistics is explanatory versus predictive modelling, both forms of supervised learning. Both serve important roles in scientific research, but a good understanding of the difference is necessary for correct usage and interpretation. Explanatory modelling is concerned with finding and testing causal theories, using data mostly as a tool. The aim here is to find an interpretable statistical model that confirms or rejects the underlying theories. Modelling becomes predictive modelling when the aim is to find the best model for predicting the outcome for unseen observations. The priority here is to generate accurate predictions, rather than being able to understand underlying associations.

Generally, predictive modela use data-driven methods, whereas explanatory models are more hypothesis-driven. There is also a difference in optimisation goal: explanatory models focus on minimizing bias (errors resulting from erroneous assumptions in the model), while minimizing estimation variance (fluctuations resulting from the specific data used) is less important since the models are used for population-level inference. Prediction models are intended to be used for individual-level predictions, and therefore the aim is to minimise estimation variance in addition to bias. In practice,

the biggest difference between explanatory and predictive modelling is the handling of data. In explanatory models, the full dataset can be used to fit the model and calculate coefficients. Validation of the model is done through goodness-of-fit tests and residual analysis to assess how well the model captures the data. In prediction models, however, the models are fit only on the training data, and a separate test dataset is used to validate the accuracy by comparing the predicted to the actual outcomes. It is important to keep the distinction between explanatory and prediction models in mind to avoid conflating the two and drawing incorrect or unsupported conclusions. [14]

In this thesis, both explanatory and prediction models are used. Chapters 3 and 4 focus on explanatory models without any predictive aspect; they relate (changes in) ferritin levels to explanatory variables that are based on biological theory. We do not attempt to minimise estimation variance, as we are interested in the average effect of the variables on ferritin levels in the whole population. The same holds for the analysis of COVID-19 antibodies in Chapter 6. Predicting any individual donor's specific ferritin or antibody level would likely not be very accurate, but we are nonetheless able to explain a considerable amount of population variance with our models. Chapter 5 is neither explanatory nor predictive, as it uses unsupervised learning methods to cluster hemoglobin trajectories and there is no outcome to explain or predict.

Chapters 7 through 9 focus on predictive modelling, which is also expressed by the term 'prediction model' in the titles. In these papers, models are trained using only the training data, and model performance is assessed using the prediction accuracy on a separate set of test data. Since predictions would be used for forecasting, the test set always contains more recent data than the training set. Although these papers focus on prediction, they are not void of explanatory aspects. Especially in Chapter 7, the explanatory aspects are just as important as the predictive aspects. The main difference with 'true' explanatory models is that Chapter 7 focuses on explaining predictions, rather than explaining causal relations with the outcome variable. Where possible, prediction explanations are related to known or hypothesised causal explanations, but this is not required to keep a predictor variable in the model.

2.2.2 Classification for donor deferral

In machine learning, classification refers to the use of a prediction model where class labels are predicted for input data. In binary classification, there are exactly two outcome classes: one is the positive outcome class and the other the negative outcome class, where the positive outcome class is typically the outcome that is of most interest to the researcher. In our context, the two outcome classes for hemoglobin deferral prediction are 'deferral' and 'non-deferral', with deferral being the positive outcome class (even though deferral is not positive in a colloquial sense).

A hemoglobin deferral prediction model could be used by blood banks in the process of inviting donors. In addition to inviting donors based on their eligibility (i.e., has the minimum donation interval been met) and current demand for different blood groups, the prediction of the model could be taken into account. By only inviting donors that are predicted to belong to class non-deferral by an accurate model, the number of donors that are deferred on-site could be reduced.

For each observation, predicting the outcome leads to one of four scenarios:

1. Prediction non-deferral, true outcome non-deferral (true negative, TN)
2. Prediction non-deferral, true outcome deferral (false negative, FN)
3. Prediction deferral, true outcome deferral (true positive, TP)
4. Prediction deferral, true outcome non-deferral (false positive, FP)

If donor invitations are guided by a prediction model, the two types of incorrect predictions (false negatives and false positives) have different effects. For false negatives, the donor is invited to donate and then deferred before donation because the hemoglobin level is below the threshold. For false positives, the donor is not invited even though their hemoglobin level would have been sufficient: a missed donation for the blood bank. A balance must be found between the false positives and the false negatives: we want to reduce the deferral rate, which is determined by the proportion of false positives, but if we predict 'deferral' too liberally, we will find many more false negatives, which means more missed donations, and a risk for the blood bank to not have sufficient blood donations.

Generally, prediction models work best when all outcome classes have the same probability of occurring. In this case, outcome class non-deferral is much more likely than outcome class deferral. Deferral rates in the Netherlands are currently at around 3% for women and 1% for men, meaning that 97% and 99% of donation attempts belong to the outcome class non-deferral, respectively. This has important consequences

for the assessment of model performance. Often, performance is measured as accuracy on a test set (accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$). Another popular choice for model performance is the ROC curve (receiver operating characteristic curve), which plots the true positive rate ($\frac{TP}{TP+FN}$) against the false positive rate ($\frac{FP}{FP+TN}$) at different classification thresholds. A classification threshold is needed to map the output of the model (the deferral probability of an observation) to a binary outcome: if the probability exceeds the threshold, the prediction will be 'deferral'. As the classification threshold decreases (i.e., a lower probability of 'deferral' is needed to classify an observation as such), more observations will be predicted as 'deferral', which increases both the false positive rate and the true positive rate. ROC curves can be summarised in one number by the ROC AUC (Area Under the ROC Curve), which is a number between zero and one, where an ROC AUC of 1 corresponds to a model with 100% correct predictions. The baseline value for the ROC AUC is 0.5: the performance of a random classifier.

However, regular accuracy and the ROC AUC may be misleading when used for datasets with imbalanced outcomes. [15] An extreme illustration of this would be a model that always predicts 'non-deferral'; although its accuracy would be 97% for women and 99% for men, such a model would not have any practical value. Similarly, the ROC curve may be too optimistic in severely imbalanced classification problems. Therefore, throughout this thesis, we use precision and recall instead of accuracy, and AUPR (Area Under the Precision-Recall curve) instead of ROC AUC. Precision is defined as the proportion of correctly predicted observations out of all observations predicted to belong to that class ($\frac{TP}{TP+FP}$ for the positive class, or $\frac{TN}{TN+FN}$ for the negative class). The precision of class deferral is the proportion of true deferrals out of all predicted deferrals. Recall is defined as the proportion of correctly predicted observations in one outcome class ($\frac{TP}{TP+FN}$ for the positive class, $\frac{TN}{TN+FP}$ for the negative class). The recall of class deferral is the proportion of deferrals that are correctly predicted to be deferrals.

There always exists a trade-off between precision and recall: by increasing the classification threshold, the number of false positives decreases, whilst the number of false negatives increases. This causes precision to increase, and recall to decrease. The graph showing precision and recall at different classification thresholds is called the precision-recall curve, and the area underneath this curve is the AUPR. Similar to the ROC AUC, AUPR is a number between zero and one, but its baseline and interpretation are different. For the AUPR, the baseline value is dependent on the proportion of observations belonging to that outcome class. In our case, the baseline AUPR for class deferral would be 0.01 for men and 0.03 for women. The AUPR should

therefore always be interpreted in combination with the baseline value.

An interesting feature is that true negatives are used for the calculation of neither precision nor recall. This is what makes it suitable for imbalanced data: we ignore the large number of non-deferrals that are predicted correctly and focus on the correctness of the prediction of deferrals, and on incorrectly predicted non-deferrals.

The aforementioned metrics provide a fair evaluation of model performance on our imbalanced dataset. All studies involving prediction models within this thesis report these metrics, or a subset thereof, depending on their relevance to the specific research question. Whenever possible, we translate these metrics into statements that clearly show their practical implications. For instance, we calculate the hypothetical deferral rate by taking the complement of the recall of class non-deferral, providing insight into the potential impact of using this model to guide donor invitations. By reporting these metrics and their interpretations, we aim to accurately describe model performance, as well as allow for straightforward interpretation of the effect of using the model for donor management.

CHAPTER

3

First results of a ferritin-based blood donor deferral policy in the Netherlands

Published in Transfusion 60(8): 1785-1792. doi:10.1111/trf.15906

Authors: M Vinkenoog, K van den Hurk, M van Kraaij, M van Leeuwen, MP Janssen

Abstract

Background - Whole blood donors are at risk of becoming iron deficient. To monitor iron stores, Sanquin implemented a new deferral policy based on ferritin levels, in addition to the traditional hemoglobin measurements.

Methods - Ferritin levels are determined in every fifth donation, as well as in all first-time donors. Donors with ferritin levels < 15 µg/L (WHO threshold) are deferred for 12 months; those ≥ 15 and ≤ 30 µg/L for 6 months. The first results were analysed and are presented here.

Results - The results show that 25% of women ($N = 20151$) and 1.6% of men ($N = 10391$) have ferritin levels ≤ 30 µg/L at their first blood center visit. For repeat (non-first-time) donors, these proportions are higher: 53% of women ($N = 28329$) and 42% of men ($N = 31089$). After a 6-month deferral, in 88% of returning women ($N = 3059$) and 99% of returning men ($N = 3736$) ferritin levels were ≥ 15 µg/L. After a 12-month deferral, in 74% of returning women ($N = 486$) and 95% of returning men ($N = 479$) ferritin levels increased to ≥ 15 µg/L.

Conclusions - Deferral of donors whose pre-donation ferritin levels were ≥ 30 µg/L might prevent donors from returning with ferritin levels < 15 µg/L. This policy is promising to mitigate effects of repeated donations on iron stores.

Introduction

Sanquin is the national blood service in The Netherlands. In addition to securing safe blood products for patients, it has a responsibility to its voluntary non-remunerated donors to diminish the risk of developing health problems related to whole blood donation. One of these risks is iron deficiency anemia or iron deficient erythropoiesis. During whole-blood donation, a donor gives half a liter of blood, containing 210 to 240 mg iron bound to hemoglobin. [16] This iron is first replaced from iron stores (of which ferritin level is an indicator), which are then slowly replenished by an increased iron uptake from food. These stores are on average 411 mg in women under 50, 591 mg in women over 50, and 880 mg in men. [16, 17, 18] Thus, the amount of iron lost during donation is relatively large in comparison to the total iron stores, especially in premenopausal women.

To monitor donors' iron statuses, donors' hemoglobin levels are measured before each donation using a photometer (HemoCue) after finger prick sampling. Donors are eligible for donation if their hemoglobin level is at least 7.8 mmol/L (12.6 g/dL) for women, or 8.4 mmol/L (13.5 g/dL) for men. A hemoglobin level below this threshold may indicate iron deficiency anemia, which needs to be prevented. Yet, donors with normal hemoglobin levels can already be iron deficient without anemia. [19] This happens when the body is not given enough time to replenish its iron stores between donations, using only hemoglobin measurements as an iron marker.

Several studies have analysed iron recovery after donation with similar results. [20, 12] In a study of 50 male donors, followed after whole blood donation, blood volume is restored first. About four days post-donation, hemoglobin levels are at the lowest point and start to increase as stored iron is released to replenish hemoglobin. At the same time, ferritin levels decline and reach their lowest point about 29 days post-donation. After 56 days (the minimum interval between two whole blood donations for men in The Netherlands), average measured ferritin levels are 27 µg/L in repeat male donors, compared to an average of 49 µg/L directly prior to donation. At that time point, the average hemoglobin level is 9.1 mmol/L, almost back to the average starting value of 9.2 mmol/L. [12] Donors in this study did not take iron supplements.

Several strategies to better monitor iron status in donors have been proposed, such as hemoglobin-guided donation intervals, ferritin-guided donation intervals, and iron supplementation. [21] Sanquin has chosen to implement a ferritin-based deferral policy for its donors. The policy started in November 2017; donors are deferred for 6 or 12 months in case their ferritin levels are ≤ 30 or < 15 µg/L respectively, even though

their hemoglobin was above the threshold and they were eligible to donate otherwise. These thresholds were based on WHO standards, which state that ferritin levels < 15 µg/L indicate iron deficiency, while higher levels reflect the size of the iron stores. [11] However, one should be aware that ferritin is also an acute-phase reactant. [8]

The main aim of the policy is to prevent donors' ferritin levels from dropping below 15 µg/L. Without regular ferritin testing, donors with low ferritin levels (≤ 30 µg/L) but hemoglobin levels above the threshold will keep donating every few months, with the risk that their iron reserves decline until hemoglobin levels fall below the threshold. By measuring ferritin levels every fifth donation, Sanquin tries to prevent donors from future deferral, thus preventing them from becoming overt iron deficient (with or without anemia). The choice to measure ferritin every fifth donation rather than at a different frequency is arbitrary and not based on extensive research.

Data on hemoglobin and ferritin levels collected during the first 18 months since the implementation of this ferritin deferral protocol were analysed to determine:

1. the distribution of ferritin levels in new donors, providing a reference distribution of ferritin levels in healthy individuals that have never donated blood before;

2. the difference in ferritin distribution between new and repeat donors;

3. the difference in donor ferritin levels before and after deferral, which provides information on the effectiveness of donor deferral to prevent donors from returning to donate with iron deficiency.

In evaluating the deferral policy based on ferritin levels, there are three important aspects to consider. The first is the effectiveness of the policy in preventing donors returning with ferritin levels below 15 µg/L. The second and third are the effects of the policy on the blood supply and on donor health, respectively. This article analyses the first aspect in depth; an exhaustive analysis of all three aspects is outside the scope of the current study and will become possible in due time.

Methods

At Sanquin, the national blood establishment in The Netherlands, every person who signs up to become a blood donor is first invited for a donor intake. This initial visit is meant to screen for infectious diseases and assessment of blood type and potential antibodies, without donation. Prospect donors that meet all the criteria of the donor

health questionnaire and have a negative infectious disease and antibody screen become a blood donor and are invited for their first donation a few weeks later.

The ferritin-based deferral policy prescribes that ferritin is measured at the intake visit for all first-time donors, and at every fifth donation in repeat donors. Donors are considered *first-time* donors only for their first donation and are considered *repeat* donors after that. Unlike hemoglobin, which is measured by point-of-care testing and gives the result directly, ferritin is measured in serum samples which are analysed within a few days after the donation has taken place. At the intake, this makes no difference, because no donation takes place during this visit. However, for repeat donors, the ferritin level is assessed after the donation has taken place, from a sample pouch that is collected along with the donated blood. This means that donors are deferred after donation (they are notified of their deferral by letter), and that ferritin measurements are available from repeat donors that have hemoglobin levels above the donation threshold only. There is currently no evidence that donating with low ferritin levels is dangerous or unhealthy, as long as hemoglobin levels are adequate. Therefore, this donation is considered to be safe even if the ferritin measurement comes back below the threshold.

Ferritin levels are assessed with the Architect i2000 by Abbott Diagnostics. Ferritin levels are divided into three categories with different consequences for the donor:

- Ferritin < 15 µg/L: the donor is deferred from donation for 12 months;
- $15 \leq$ Ferritin ≤ 30 µg/L: the donor is deferred from donation for 6 months;
- Ferritin > 30 µg/L: no deferral, the donor can return for the next donation after the regular minimum donation interval (56 days for men, 122 days for women).

Sanquin does not have a policy to advise donors to take iron supplements for low ferritin or hemoglobin levels, although they are free to take over-the-counter iron supplements on their own initiative. The deferral periods are meant to give the donors a break from blood donation, allowing iron stores to recover solely by iron intake from donors' regular diets.

Sanquin collects approximately 400,000 whole blood donations annually, from over 270,000 donors. [22] Data for this study were collected between November 2017 and April 2019 on donors who gave consent for the use of their data for scientific research (more than 99% of all donors give this consent).

To compare the ferritin distributions in first-time and repeat donors, for each donor only the first ferritin measurement is considered. For first-time donors, this is

the ferritin measurement taken at the pre-donation screening. For repeat donors, this is the ferritin measurement taken at the fifth donation since the implementation of the protocol. If the same donor has a consecutive ferritin measurement, five donations later, that measurement is not used in this analysis, so that every donor only occurs once in the data set.

To assess the effectiveness of the deferral for preventing donors from returning to donate while iron deficient, we compare pre-deferral ferritin levels to post-deferral ferritin levels, of all deferred donors of whom post-deferral ferritin measurements were available. We compared pre-deferral ferritin levels in donors with and without a post-deferral measurement to check for selection bias. In donors without post-deferral measurement, we selected only those who were eligible for donation again (i.e., their deferral period has ended).

For donors who do have a post-deferral ferritin measurement, we calculated the average daily increase in ferritin levels for each donor. Note that since ferritin recovery does not progress linearly, the averages do not represent the actual increase on any given day, but this method can be used to compare recovery rates between women and men. [12]

All analyses are performed in the R programming language and environment for statistical computing. Plots are produced with the ggplot2 package. Distributions are asymmetric and are therefore characterised by the median value and the interquartile range (IQR). Density plots presented are kernel density estimates; the bandwidth is selected by Silverman's rule of thumb.

Results

Ferritin levels were measured at least once in 30 542 first-time donors (20 151 women) and 59 418 repeat donors (28 329 women). Figure 3.1 shows the distribution of ferritin levels for various combinations of sex and age categories. In first-time donors, men had substantially higher ferritin levels than women, and ferritin levels increased with age: median ferritin levels ranged from 96 to 173 µg/L in men and from 43 to 81 µg/L in women by age group. In repeat donors, the median values were more similar for both sexes, ranging from 22 to 35 µg/L in men and from 28 to 36 µg/L in women. Table 3.1 shows the median ferritin level and IQR for all age groups.

Overall, 25% of female first-time donors (95% CI 24%-25%) and 1.6% of male first-time donors (95% CI 1.4%-1.8%) had ferritin levels below the threshold of 30 µg/L at the intake visit. These proportions were considerably higher in repeat donors: 53%

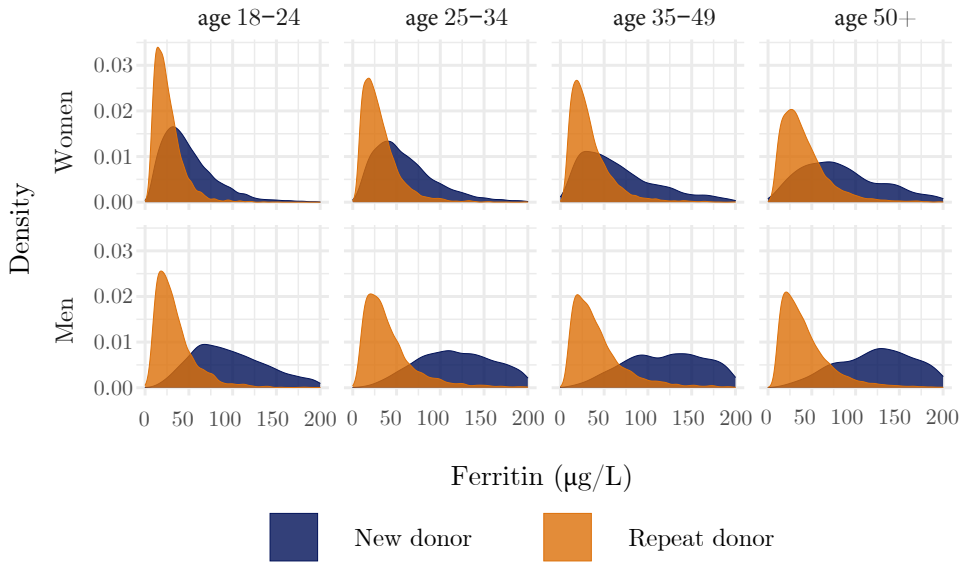

Figure 3.1: Distributions of ferritin levels in new donors (blue) and repeat donors (orange) for various combinations of sex and age category. Sample sizes range from 1037 (male new donors age 50 and up) to 14848 (male repeat donors age 50 and up).

		First-time donors		Repeat donors	
Sex	Ages	N	Median ferritin (IQR)	N	Median ferritin (IQR)
Women	18-24	9713	43 (27-65)	4537	22 (15-33)
	25-34	5071	52 (33-79)	5045	26 (17-39)
	35-49	3801	58 (33-98)	7411	28 (18-43)
	50+	1566	81 (50-128)	11,336	35 (23-53)
Men	18-24	3896	96 (66-135)	2048	28 (18-43)
	25-34	3424	136 (95-191)	3928	34 (21-53)
	35-49	2167	154 (102-224)	7063	35 (22-56)
	50+	904	173 (120-256)	18,050	36 (23-56)

Table 3.1: Median ferritin levels and interquartile range (IQR) in first-time and repeat donors by sex and age category.

of women (95% CI 52%-54%) and 42% of men (95% CI 41%-43%) had a ferritin level $\leq$ 30 µg/L. These outcomes again show that men have significantly higher ferritin levels than women (as witnessed by the confidence intervals), and that repeat donors are much more likely to have low ferritin levels than first-time donors, although this difference is much more pronounced in men (25-fold increase) than in women (two-fold increase). This leads to substantially smaller differences in ferritin levels between men and women for repeat donors than in first-time donors.

Most donors with low ferritin levels had a ferritin level between 15 and 30 µg/L, but 5.3% of female first-time donors and 0.1% of male first-time donors already had ferritin levels $<$ 15 µg/L. In repeat donors, these low levels were observed in 15% of female and 9.4% of male donors.

We calculated the moving average (window size of 1000 observations) of the proportion of donors that were deferred due to low ferritin levels as a function of age. We did this separately for sex, donor type (new/repeat donor), and ferritin deferral category ($<$ 15 µg/L and between 15-30 µg/L). In Figure 3.2, the proportion of deferrals as a function of donor age is shown for each combination of deferral type, donor type, and sex. Confidence intervals are not shown due to the proximity of the lines, but they are all extremely narrow. The difference in deferral probability between female and male donors was substantially larger in new donors than in repeat donors. In male repeat donors, the association between age and deferral rate was negative and almost linear. In female repeat donors, there was a clear non-linear dependency on age: after an initial decrease until the age of 25, there was an increase until the age of 40, after which it started to decrease again.

We also analysed the difference between pre- and post-deferral ferritin levels for both 6- and 12-month deferral. To check for selection bias, we compared the pre-deferral ferritin levels of donors with and without a post-deferral measurement. Table 2 shows the number of deferred donors, the number of donors whose deferral period has ended, and those who have already returned. It shows that pre-deferral ferritin levels in donors who returned after deferral do not differ from those in the complete group. This indicates that the group of donors with a post-deferral ferritin measurement are likely to be a representative sample of all deferred donors with respect to ferritin. However, there is a difference in return rate between the sexes: approximately 80% of men versus only 60% of women have returned out of those whose deferral period has ended. The return rates include donors who have returned after deferral but did not have a repeat ferritin measurement due to a low hemoglobin level.

After a ferritin deferral period, the deferral rate due to low hemoglobin levels is

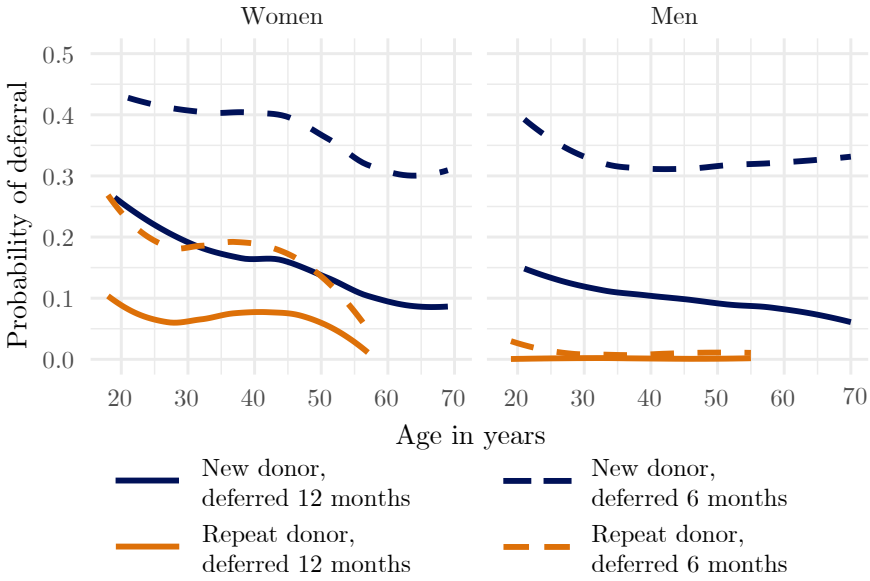

Figure 3.2: Probability of deferral due to low ferritin as a function of age in new donors (blue line) and repeat donors (orange line). Both deferral for 6 months (ferritin between 15 and 30 µg/L, dashed line) and deferral for 12 months (ferritin under 15, solid line) are shown. The difference in age ranges is due to the fact that new donors are only accepted until the age of 65, whereas repeat donors can keep donating for several more years.

	Six-month deferral		Twelve-month deferral	
	Women	Men	Women	Men
Number of donors deferred	15008	10296	5974	2952
Median ferritin at deferral (IQR)	22 (19-26)	22 (18-26)	11 (9-13)	12 (10-13)
Number of donors whose deferral period has ended	6181	4576	906	596
Median ferritin at deferral (IQR)	23 (19-26)	22 (18-26)	10 (8-12)	12 (10-13)
Number of donors returned after deferral	3258 (53%)	3883 (85%)	540 (60%)	490 (82%)
Median ferritin at deferral (IQR)	22 (18-26)	22 (18-26)	11 (9-13)	12 (10-13)

Table 3.2: The total number of donors deferred, those that are eligible to return for donation at the time of analysis for this study (deferred at least 7 months ago for 6-month deferral, and at least 13 months ago for 12-month deferral), and those that have already returned for donation. For each group the median ferritin level and interquartile range (IQR) at deferral are given in µg/L. Percentages behind the number of donors returned after deferral are with respect to the number of donors whose deferral period has ended and therefore could have returned after deferral.

considerably lower than it is in general. The overall hemoglobin deferral rate is 8.4% for women and 4.6% for men. After a 12-month deferral, 6.1% of women and 1.6% of men are immediately deferred again because their hemoglobin levels are below the threshold. After a six-month deferral, these percentages are 4.4% for women and 2.8% for men.

The changes in ferritin levels after 6- and 12-month deferrals are summarised in Table 3.3. After either deferral period, the majority of donors who returned had an increased ferritin level, men more so than women. More than 95% of returning male donors had a ferritin level of 15 µg/L or higher after either deferral type. In female donors, this proportion was 88% after six-month deferral and 73% after 12-month deferral. The difference in ferritin recovery rate between men and women makes sense considering the differences in ferritin levels observed in first-time donors. These differences can likely be attributed to the same physiological cause(s).

The rate of ferritin recovery differed between female and male donors. The median of the average daily increase in women was higher for 12-month deferral than for six-month deferral: 0.030 µg/L/day versus 0.016 µg/L/day. In men, they were more similar: 0.068 µg/L/day for 12-month deferral and 0.071 µg/L/day for six-month

	Six-month deferral		Twelve-month deferral	
	Women	Men	Women	Men
Number of donors returned	3059	3736	486	479
Donors with increased ferritin	61%	91%	91%	99%
Median total increase (IQR)	4 (-3-11)	15 (7-25)	12 (5-22)	27 (16-39)
Median increase per day	0.016	0.071	0.030	0.068
Ferritin after deferral <15 µg/L	12% (↓)	1.5% (↓)	26% (=)	4.6% (=)
Ferritin after deferral 15 - 30 µg/L	54% (=)	30% (=)	43% (↑)	27% (↑)
Ferritin after deferral >30 µg/L	34% (↑)	68% (↑)	30% (↑↑)	68% (↑↑)

Table 3.3: Ferritin levels of donors who return after 6-month deferral (ferritin level between 15 and 30 µg/L) or 12-month deferral (ferritin level < 15 µg/L). Symbols in the bottom three rows indicate whether the ferritin level has dropped (↓), has gone up one (↑) or two (↑↑) categories, or has stayed in the same category (=)

deferral. After either period of deferral, ferritin recovery rates were substantially higher in men than in women.

Discussion

This study shows that in first-time donors who have never donated blood, women's ferritin levels are lower than men's, and they increase with age. Ferritin levels in repeat donors are substantially lower and therefore the deferral rate is higher, for both sexes. The difference in ferritin levels between male and female donors is considerably smaller in repeat than in first-time donors, regardless of age. Finally, after having a measured ferritin level below 30 µg/L and being deferred for 6 or 12 months, the vast majority of returning female and almost all returning male donors have ferritin levels of 15 µg/L or higher.

The differences we have observed between male and female first-time donors can partly be explained by the effect of the menstrual cycle on iron stores. After menopause, this additional iron loss is no longer present and women's ferritin levels increase. [23] The fact that sex differences are much smaller among repeat donors suggests that regular blood donation leads to a lower ferritin level, which impacts men more than women as their natural ferritin stores are generally higher. Multiple studies have found that an increase in the number of donations results in decreased iron stores, even though hemoglobin levels remain above the threshold for donation. [24, 25] Our results suggest that this relationship is less strong in women. This can be explained by the shorter minimum donation interval for men (56 days vs. 122 days for women), which allows

them to donate five times a year, compared to three times a year for women. Also, donation frequency is the best predictor for decreased iron stores. [26] Further research into the precise relationship between donation frequency, total number of donations and trends in ferritin levels is ongoing, for instance in the INTERVAL study. [25] Another explanation is that women are more easily deferred than men; the hemoglobin threshold for donation is much closer to the average hemoglobin value in women than in men. Women with low iron stores are already deferred by the hemoglobin test alone, so their (likely low) ferritin levels have not been measured in this study.

Sex differences can also be seen in the percentage of donors that return after donation: men are more likely to return than women. Before we try to explain this difference, we should keep in mind that donors only come back after they are actively invited by Sanquin by means of an invitation algorithm (based on daily demand for blood and blood types). The effect of this procedure may hinder the outcome of the current analysis. However, studies on donor return rates after deferral are consistent in finding a higher return rate for men than for women, although the magnitude of the difference varies. [27, 28, 29]

Regarding the increase in ferritin after deferral, we assume that this is larger than what would have occurred in case the donors had not been deferred according to the policy. This assumption is based on studies mentioned in the introduction, which show that donors need at least 168 days for ferritin levels to recover. [20, 12] This indicates that a longer deferral period gives donors more time to restore their ferritin stores by taking a break from their regular donation schedule. Nonetheless, a considerable number of donors is deferred again based on their ferritin level upon their return, especially women.

In male donors, the average daily ferritin increase is higher in donors who were deferred for 6 months than in those deferred for 12 months. This is interesting, because during the first 29 days after donation, ferritin levels are still decreasing. [12] It might indicate that after the initial decrease, ferritin recovery starts off at a high rate which then tapers off. We see different results in women: the average daily ferritin increase is higher in women deferred for 12 months than those deferred for 6 months. Even though recovery rates of donors deferred for 12 versus 6 months cannot be compared because of their different ferritin levels before deferral ($<$ 15 vs. between 15-30 µg/L), it is remarkable that the ratio between these rates differs between men and women. One explanation might be that ferritin recovery takes longer for women, so the increase starts later in the process. However, no differences between men and women were found in ferritin recovery speed in control groups of oral iron

supplementation studies, although sample sizes were relatively low (about 20 people per group). [20, 30] A larger-scale study that measures donors' ferritin levels in the weeks following a donation could provide more insight.

Some blood services supply blood donors with iron supplements in order to prevent iron deficiency, which can lead to restless-leg syndrome and pica, especially pagophagia, the inclination to chew ice. [31, 4] There is no solid evidence for an association between low iron stores and fatigue and cognition. [4] Some studies did find that iron supplements improve cognition in adolescents and women, but most of these have small sample sizes and are methodologically weak, with evidence of publication bias. [30] The INTERVAL study did not find any effect of shortening the donation interval on cognitive function in an analysis of health survey questionnaires given to more than 45,000 donors. [25] An analysis on more than 16,000 donors participating in the Danish Blood Donor Study did not find an association between low ferritin levels and self-reported mental and physical health either. [32]

Regardless of its possible health effects, several studies have shown that iron supplementation increases the speed of recovery of iron stores and hemoglobin levels after blood donation. [20, 33] However, iron supplementation can also have unintended and unwanted side effects, which may impact compliance of iron supplementation and can deter donors. For this reason, as well as the lack of scientific consensus on how iron supplementation in blood donors should be installed, Sanquin chose to introduce ferritin-guided donation intervals rather than iron supplementation to mitigate effects of repeated donation on iron stores.

Although the ferritin-guided deferral policy seems to help donors maintain appropriate ferritin levels, it also raises some concerns. In the past few years, the proportion of new female donors under 25 years of age has been increasing rapidly in The Netherlands. [34] Ferritin levels below 30 µg/L are very common among young women who have never donated blood. If this trend continues, the proportion of first-time donors that immediately gets deferred from donation based on ferritin levels will continue to increase. Deferral of these potential donors may lead to a lower availability of blood products and has a larger effect on the blood supply than hemoglobin-based deferral. One to three donations are lost for every 6-month deferral, and three to five for every 12-month deferral, depending on sex. Additionally, by deferring donors not only for low hemoglobin, but also for low ferritin levels, the chances of a donor being deferred are increased. However, for the long-term this increased chance may decrease again, as deferral due to low ferritin can lower hemoglobin deferral rates. In our data set, we found that the hemoglobin deferral rate decreases by half after ferritin deferral as

compared to the overall hemoglobin deferral rate. Deferral can also cause donors to become unmotivated and not return to the blood center, especially first-time donors. [27, 28] Compensating for lost donations by recruiting new donors could therefore be a less desirable consequence. Therefore, it is important to carefully monitor donor availability when introducing ferritin-guided donation intervals. One should also note that the frequency of measuring ferritin levels (every fifth donation) is mostly arbitrary and loosely based on a trade-off between cost and benefit. Measuring more often would identify donors at risk of iron deficiency earlier, but also increases cost and loss of potential donations due to deferral.

From our results, we conclude that repeat donors have considerably lower ferritin levels and smaller differences between sexes in comparison to first-time donors. Deferral of donors with ferritin levels ≤ 30 µg/L seems to prevent the majority of donors, male donors in particular, from returning to donate with iron deficiency.

Comparisons to a control group are needed to establish whether ferritin levels are indeed higher in groups of donors than they would have been without ferritin-guided donation intervals. Furthermore, longer-term research is needed to assess whether this policy can maintain donors' ferritin levels within the appropriate range.

CHAPTER

4

Individual and environmental determinants of serum ferritin levels: a structural equation model

Published in Transfusion Medicine 33(2): 113-122. doi:10.1111/tme.12902

Authors: M Vinkenoog, R de Groot, J Lakerveld, MP Janssen, K van den Hurk

Abstract

Background - Serum ferritin levels are increasingly being used to assess iron stores. Considerable variation in ferritin levels within and between individuals has been observed, but our current understanding of factors that explain this variation is far from complete. We aim to combine multiple potential determinants in an integrative model, and investigate their relative importance and potential interactions.

Methods - We use ferritin measurements collected by Sanquin Blood Bank on both prospective ($N = 59596$) and active blood donors ($N = 78318$) to fit a structural equation model with three latent constructs (individual characteristics, donation history, and environmental factors). Parameters were estimated separately by sex and donor status.

Results - The model explained 25% of ferritin variance in prospective donors, and 40% in active donors. Individual characteristics and donation history were the most important determinants of ferritin levels in active donors. The association between environmental factors and ferritin was smaller but still substantial; higher exposure to air pollution was associated with higher ferritin levels, and this association was considerably stronger for active blood donors than for prospective donors.

Conclusions - In active donors, individual characteristics explain 20% (17%) of ferritin variation, donation history explains 14% (25%) and environmental factors explain 5% (4%) for women (men). Our model presents known ferritin determinants in a broader perspective, allowing for comparison with other determinants as well as between new and active donors, or between men and women.

Introduction

Iron is essential for human life, but both iron deficiency and iron overload can cause various adverse health effects. Therefore, iron homeostasis is tightly regulated in humans. In case of insufficient availability of iron in the circulation, recycling of old red blood cells is increased and hepcidin is downregulated both to increase dietary iron absorption and release iron stored in ferritin. [8, 35] Hemoglobin levels have long been the standard method to assess iron status. However, hemoglobin levels can remain sufficient for some time, even when iron stores are dwindling; this is known as iron deficiency non-anemia. [8]

In contrast to hemoglobin, serum ferritin levels reflect the amount of stored iron. [8] Therefore, they are increasingly used to assess individuals' iron stores when these are at risk, for instance after traumatic blood loss, during pregnancy, or in blood donors. [21] Sanquin, the national blood service in the Netherlands, started measuring ferritin levels in each new donor, and subsequently after every fifth donation, in October 2017. Donating blood has a substantial impact on ferritin levels. Ferritin levels are lower among blood donors than in the general population: cross-sectional studies report lower ferritin levels in donors with a higher number of whole blood donations and a large randomised trial showed that ferritin levels indeed decline with more frequent blood donations. [36, 37] Among new donors, large variation in ferritin levels is observed. [36] It is well established that individual characteristics such as sex and age are relevant: women in general, but pre-menopausal women in particular, have considerably lower ferritin levels than men. [36, 2, 38] Higher body mass index (BMI) is associated with higher ferritin levels. [39] In recent decades, many other factors that affect iron status have been identified: diet, [40, 41] genetics, [42, 43] ethnicity, [44] and iron supplementation, which is mostly studied among blood donors. [20, 45]

Ferritin is also a known acute-phase protein that is elevated in inflammatory conditions, complicating its diagnostic value in individuals with conditions such as inflammatory bowel disease or chronic heart failure. [9] This could also explain the association between BMI and ferritin levels, as adipose tissue is known to promote systemic inflammation. [46] Additionally, exposure to environmental pollutants has been linked to disordered iron homeostasis, [47, 48] and ambient particle matter (PM) concentration is correlated with ferritin levels. [48] The biological mechanism behind this is still unclear, but it is postulated that iron attaches to the PM rather than to cell nuclei, effectively creating a functional deficiency. [47, 48] In turn, mechanisms start upregulating iron uptake and recycling in an attempt to meet the iron require-

ment of the cells, thereby altering iron homeostasis. Another suggested mechanism is that when pollutants enter the lungs, iron is transported away from the surface of the lung tissue and stored in ferritin complexes, in order to avoid chemical reactions between iron and the pollutant. [47] Other potential environmental determinants are neighbourhood characteristics, including population density and socio-economic status, which are consistently shown to be related to body weight and blood parameters. [49]

Previous studies on ferritin levels have focused on studying the association with variables in a limited setting, for example, characteristics such as age and BMI, donation-related variables, or environmental pollutants. In this paper, we propose a novel framework that integrates multiple settings, using a structural equation model. By grouping relevant explanatory variables into constructs, we describe relationships with ferritin on a more general level. This enhances the insight into various mechanisms that influence ferritin levels, which is valuable to those who use these as a diagnostic tool. We explore associations between ferritin levels and individual characteristics, donation behaviour and environmental factors, in a large group of newly registered and active whole blood donors.

Methods

For this cross-sectional study, data collected by Sanquin and the Geoscience and health cohort consortium (GECCO) were analysed. Sanquin is by law the only blood service in the Netherlands, collecting over 400 000 whole-blood donations each year, with collection sites geographically well-distributed throughout the country. Several eligibility criteria exist to ensure the safety of the donors and recipients and the quality of the blood product. Donors must be aged between 18 and 79 years old, and a pre-donation screening visit takes place before the first 500 mL whole blood donation, which includes blood sampling for blood type and infectious disease testing, as well as initial hemoglobin and ferritin measurements. We will refer to these prospective donors, who have not donated yet, as *new donors*.

Before every donation, a donor screening is performed, including a donor health questionnaire and measurements of blood pressure, pulse rate and hemoglobin levels to assess whether the donor is eligible to donate. Hemoglobin levels need to be at least 7.8 mmol/L for women and 8.4 mmol/L for men. This is measured by point-of-care testing with a photometer (HemoCue, Angelholm, Sweden). Ferritin levels, are measured in serum samples, using the Architect i2000 (Abbott Diagnostics, Chicago,

IL), after the pre-donation screening visit and after every fifth whole blood donation. As such, ferritin measurements are only available in case of successful whole blood donations, and for new donors whose venous samples are taken as part of the pre-donation screening visit.

Data

This study included all new and active whole blood donors who gave consent to the use of their data for scientific research (this consent is given by > 99% of all donors) and for whom ferritin measurements were available between 1 October 2017 and 31 December 2019. If multiple ferritin measurements were available for a donor, only the first measurement was used. Information on donors and donation histories was extracted from the blood bank information system (ePROGESA, MAK-SYSTEM International Group, Paris, France). Variables used were sex, age, height, weight, time since previous successful donation, the number of successful donations in the previous 2 years, donor status (new or active donor), and ferritin levels. BMI was calculated from self-reported donor height and weight. Sanquin does not register donor ethnicity, but Duffy negative phenotype was included to function as a proxy for sub-Saharan African descent.

Environmental exposure variables of various characteristics were obtained from the Geoscience and health cohort consortium (GECCO). [50] The exposure data were operationalised based on publicly available data. Data from 30 weather stations in the Netherlands—obtained from the Royal Netherlands Meteorological Institute (KNMI)—were used to estimate temperature at a spatial resolution of 1 km. Three options for the measurement level were considered (minimum, average, and maximum daily temperature), as well as three time spans (day, week or month before donation), resulting in nine options in total. The combination that showed the highest correlation with ferritin was included in the final model.

Daily concentrations for particulate matter (PM) 2.5, PM10, NO2, ozone and soot levels were obtained via the Dutch National Institute for Public Health and the Environment (RIVM), for the years 2017–2019. These variables were imputed on a spatial resolution of 1 by 1 km. Neighbourhood socio-economic status (SES) scores and population density from 2017–2019 were acquired from Statistics Netherlands (CBS), both available on 6-digit postal code level. SES scores are based on percentiles of income, education level and vocational history of households, with a score of 0 being exactly the national average, and positive scores being above average. All spatio-temporal variables were matched with donor and donation data based on donation date and donor postal code. Lastly, the date and time of each donation were included

as potential factors to account for seasonal and diurnal variation, as they are known to affect hemoglobin levels and may also affect ferritin levels.

To check for a possible confounding effect of smoking on environmental variables, we analysed the correlation between the percentage of smokers per municipality (data from Statistics Netherlands) and all environmental variables described in the above paragraph.

There were no missing data for environmental datasets from the RIVM and CBS. Donors with no ferritin measurement were excluded from the analysis. There were no missing data for the other donor or donation level variables.

Statistical analysis

Structural equation modelling (SEM) was used to investigate which variables relate to serum ferritin and to what extent. Briefly, observed variables and latent constructs are distinguished in SEM. Latent constructs cannot be measured or observed directly, but are inferred from the observed variables. One or more hypothesised sets of relationships and correlations between variables and constructs are specified a priori and shown in a path diagram. For each relationship, a parameter is estimated that indicates its strength. Estimates are obtained by numeric optimisation of a fit criterion, using maximum likelihood estimation. A more detailed overview of this method is provided in the Appendix.

We compared four ways to divide the 15 variables included in the analysis into latent constructs, as shown in Table 4.1. Date and time of the donation were added to the model separate of the constructs, and as such are not included in Table 4.1. Model A contains four latent constructs, and in models B, C and D different sets of constructs are combined. Confirmatory factor analysis (CFA) was used to test the validity of the specified measurement models, that is, the hypothesised relationships between the latent constructs and their observed variables. The overall fit of the models was assessed by the Tucker-Lewis Index (TLI) and the root mean square error of approximation (RMSEA). A rule of thumb is to exclude variables for which the absolute value of the standardised factor loading is below 0.4, but at sample sizes larger than 300, if the overall model fit is good, exclusion is not necessary and should be judged separately for each variable based on sensible background knowledge. [51]

Pairwise residual correlations between observed variables were calculated to identify whether any covariances needed to be added to the model. Of the four specified models, we continued our analysis with the best fit according to CFA, based on the TLI and RMSEA.

Variable	Model A	Model B	Model C	Model D
Age Height Weight BMI Duffy phenotype	Individual characteristics	Individual characteristics	Individual characteristics	Individual characteristics
Time since prev. donation Number of prev. donations	Donation history	Donation history		
Population density Temperature Socio-economic status	Environment	Environment	Environment	Environment
Ozone PM2.5 PM10 Soot NO2	Pollution		Pollution	

Table 4.1: Grouping of variables into constructs for each model. Note that variables *time since previous donation* and *number of previous donations* are only available for active donors.

To the model with the best fit, we added the structural component, which contains the relationships between the latent variables and ferritin, the outcome variable. A multiple group SEM was carried out with parameters estimated separately for male and female donors, and for new and active donors. Because the assumption of normality of the explanatory variables does not hold in our data, a different estimator than the default maximum likelihood estimator was used: the 'mean and covariance adjusted weighted least squares estimator', which is robust against violations of the normality assumptions in a multivariate setting. [52]

The same model was fitted in all four groups, although the variables belonging to the donation history construct (see Table 4.1) are not available for new donors, as they do not (yet) have a donation history. The overall fit of the SEM model was assessed using the TLI and RMSEA, as well as the R2 measure.

All analyses were conducted using R programming language and environment for statistical computing version 4.0.3, with package zoo for pre-processing environmental data, and lavaan for CFA and SEM analyses. Path diagrams were created with yEd Live Graph Editor.

Results

Sample composition

Table 4.2 shows descriptive statistics of the study population by sex and donor status. The size of each of the groups was comparable, except for the group of new male donors, which was only half the size of the other groups. Between new and active donors, age differed considerably, new donors being younger than active donors by 17 years on average ($p < 0.001$). In both new and active donors, men were older (by 6 years on average, $p < 0.001$) and heavier (by 13 kg on average, $p < 0.001$) than women. P-values were obtained using two-sample t-tests. The time since last donation is higher in women than in men, and the number of prior donations is higher in men than in women. These differences are due to differences in the minimum required donation interval: for women, there must be 122 days between two donations with a maximum of 3 donations per year, while for men, the minimum is 57 days between two donations with a maximum of 5 donations per year. Differences in ferritin levels between the groups are as expected from previous studies: men have higher ferritin levels than women, and repeat donors have lower ferritin levels than new donors.

For pollution and environmental variables, there was little difference between the

groups, any differences between new and active donors were most likely due to the different age and geographical distribution of the groups. None of these differences were statistically significant.

We found a weak correlation between the percentage of smokers and SES score (Pearson's $r = -0.4$) and a moderate correlation between the percentage of smokers and population density (Pearson's $r = 0.5$). No correlation was found for any of the other environmental variables.

	New donors		Active donors	
	Women	Men	Women	Men
N	40172	19424	39085	39233
Age (years)	26 (21–37)	28 (23–37)	47 (31–58)	53 (39–62)
Height (cm)	170 (166–175)	183 (178–188)	170 (166–175)	183 (178–188)
Weight (kg)	68 (62–77)	82 (74–90)	70 (64–80)	85 (78–93)
BMI (kg/m2)	24 (21–26)	24 (22–27)	24 (22–27)	25 (23–27)
Time since prev. donation (days)	NA	NA	154 (132–217)	139 (71–147)
Number of prev. donations	NA	NA	3 (2–4)	5 (4–7)
Population density (per km2)	1173 (425–2617)	1246 (477–2936)	827 (322–1855)	814 (320–1824)
Duffy phenotype (proportion)	0.25	0.17	0.28	0.16
Temperature (°C)	11.4 (6.4–16.6)	11.7 (6.6–16.7)	10.4 (6.0–16.0)	10.4 (5.9–16.0)
Socio-economic status	0.04 (-0.21 to 0.22)	0.02 (-0.24 to 0.22)	0.10 (-0.10 to 0.25)	0.12 (-0.07 to 0.26)
Ozone (µg/m3)	46.9 (45.6–48.8)	46.8 (45.5–48.7)	47.2 (45.9–49.2)	47.2 (45.9–49.1)
PM2.5 (µg/m3)	10.7 (9.7-11.6)	10.7 (9.8–11.6)	10.5 (9.6–11.5)	10.6 (9.7–11.6)
PM10 (µg/m3)	18.2 (16.8–19.3)	18.2 (16.9–19.3)	18.0 (16.6–19.0)	18.0 (16.7–19.1)
Soot (µg/m3)	0.66 (0.54–0.78)	0.66 (0.55–0.78)	0.63 (0.52–0.75)	0.65 (0.54–0.76)
NO2 (µg/m3)	17.6 (14.9–21.6)	17.8 (15.1–21.8)	16.8 (14.2–19.7)	16.9 (14.3–19.6)
Ferritin (µg/L)	47 (28–75)	118 (79–170)	30 (17–47)	34 (20–56)

Table 4.2: Distribution of explanatory variables by donor status and sex.

Model selection

CFA did not provide support for the environment construct as defined by the three variables temperature, population density and socio-economic status. These variables did not share a high proportion of their variance and consequently there was no convergent validity, effectively ruling out models A and C. In models B and D, variables Duffy phenotype, temperature, SES and height were omitted due to very low factor loadings (< 0.05). The factor loading for variable age was also low (0.35) but this variable was not excluded, as it is expected that this factor loading would be small, considering the other variables in the construct (weight and BMI) are much more closely related. All other factor loadings were above the suggested threshold of 0.6. All latent constructs (individual characteristics, donation history and environment) showed convergent and discriminant validity in models B and D. Variables time and day of year, which were added to the model outside the constructs, were also dropped due to very low factor loadings (< 0.05).

The presence of a donation history construct was the only difference between models B and D, and since new donors do not yet have a donation history, the models only differed for active donors. Model B had a TLI of 0.961 and RMSEA of 0.063, while model D had a TLI of 0.932 and RMSEA of 0.083. Based on these fit measures, model B fit the data best, and was therefore used in the remainder of the analyses.

Based on inspection of the pairwise residual correlations between all observed variables, two covariance terms were added to the model: one for PM2.5 and PM10 (residual correlation 0.092 to 0.102, depending on sex/donor status), and one for age and population density (residual correlation -0.151 to -0.149, depending on sex/donor status). We also added one covariance term for weight and BMI, as BMI was calculated using weight and was therefore inherently dependent.

Parameter estimates

Figure 4.1 shows the structure of the final model and the parameter estimates for new donors. Parameter estimates were similar for both sexes, but factor loadings for variables belonging to the individual characteristics construct were higher for women than for men, indicating more shared variance. Factor loadings in the environment construct did not differ between sexes, showing that the covariance structure of those variables was not dependent on sex. The parameter estimates for the regression coefficients show the relative importance of each latent construct for the outcome variable. Table 4.3 shows the percentage of variance in ferritin levels that is explained by each

Construct	New donors		Active donors	
	Women	Men	Women	Men
Individual characteristics	23%	23%	20%	17%
Donation history	NA	NA	14%	25%
Environment	2%	2%	5%	4%
Total variance explained	25%	25%	39%	46%

Table 4.3: Relative contribution to explanation of variance of ferritin levels per model.

construct for each model, adding up to the total percentage of variance explained.

Figure 4.2 shows the final model for active donors. As in new donors, factor loadings in the individual characteristics construct were higher for women than for men, and they were also higher for new donors than for active donors. The relative importance of individual characteristics and donation history was opposite for both sexes: for men, donation history was correlated with ferritin levels more strongly than individual characteristics (0.66 vs. 0.45), while this was reversed for women (0.43 vs. 0.61). The regression coefficient of the environment construct is 0.15 for women and 0.10 for men. The environment construct explains twice as much variation in ferritin levels in active donors as in new donors.

As for overall model fit, with a TLI of 0.981 and 0.979 and RMSEA of 0.052 and 0.042, for new and active donors respectively, both models fit very well when compared to commonly used thresholds (TLI > 0.95, RMSEA < 0.06). [53] R2 was calculated separately by sex: for new donors, R2 was 0.251 for men and 0.252 for women, and for active donors, 0.458 for men and 0.393 for women.

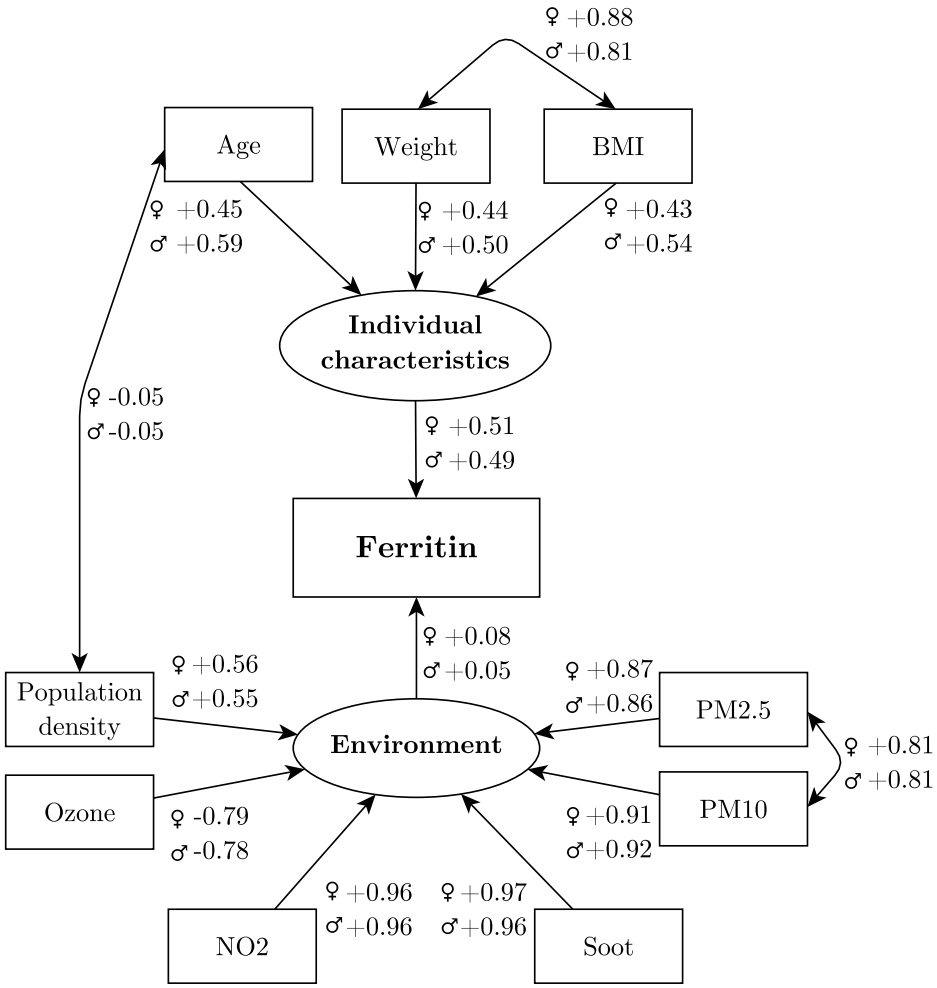

Figure 4.1: Final structural equation model for ferritin determinants in new donors, with parameters estimated separately for men and women. All parameter estimates are standardised so that the variance of each observed variable and latent construct equals 1.

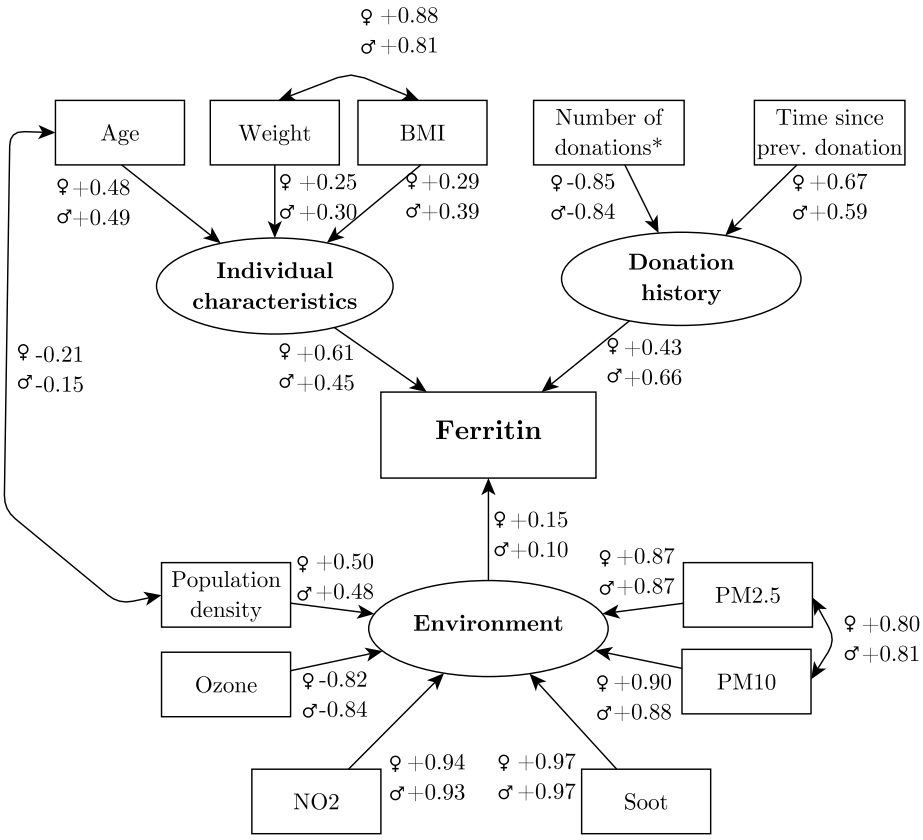

Figure 4.2: Final structural equation model for ferritin determinants in active donors, with parameters estimated separately for men and women. All parameter estimates were standardised so that the variance of each observed variable and latent construct equals 1.

Discussion

This study investigated the impact of individual and environmental determinants on ferritin levels in Dutch individuals, using SEM. The model was able to explain 25% of ferritin level variance in new donors for both sexes, and 46% and 39% in active donors for male and female donors, respectively.

We found the construct composed of individual characteristics (age, weight, and BMI) to be the most important determinant of ferritin in female active donors, followed by donation history (time since previous donation, number of donations in the past 2 years). For male active donors, this was the opposite: donation history was a more important determinant than individual characteristics. In both sexes, environmental factors are associated with ferritin levels, albeit to a lesser degree than individual characteristics and donation history.

The relationship between ferritin levels and anthropometric characteristics is well-documented, and the positive correlations we found for ferritin with age, weight and BMI are consistent with those found in other studies. [36, 45, 54] Men have much higher ferritin levels than women in general and show a larger decrease in ferritin levels after repeated donations. As a result, ferritin levels in active donors are similarly low for women and men. [36] The donation history construct explained more variance in ferritin levels in men than in women. Although often not explicitly mentioned, this discrepancy is also found in previous studies, with stronger relationships between variables regarding donation history and ferritin for men than for women. [45] A reasonable explanation for this is that men commonly display more variation in donation history variables due to the possibility of more frequent donations: in many blood services, men are allowed to donate more often than women and are usually less frequently deferred for low hemoglobin levels. [55]

From previous epidemiological studies, we know that environmental factors may play a role in iron metabolism, and that certain pollutants can disrupt iron homeostasis. [56] Our study shows that although environmental factors are less strongly associated with ferritin levels than individual characteristics and donation history, their effects are far from negligible. Because of the wide reach of environmental exposures over geographic areas, even a relatively small influence on individuals can result in a large effect on the population level. As this study includes only data from the Netherlands, which is a relatively small country, associations between environmental variables and ferritin levels were not very strong, as was expected. Repeating this study on a larger, or even global, scale may result in finding a more substantial effect.

Higher values for all but one environmental factor (ozone) were positively correlated with higher ferritin levels. These findings support the hypothesis that air pollution causes higher ferritin levels. The underlying mechanism may be that when certain pollutants enter the lungs, iron is transported away from the lung tissue surface and stored in ferritin complexes to avoid chemical reactions between iron and the pollutant. [47, 57] This would imply that using serum ferritin as a proxy for total body iron is less reliable when there is significant air pollution.

The environment construct was more strongly associated with ferritin level in active donors than in new donors. In new donors, environmental factors explain 2% of variance in ferritin levels, while in active donors this increases to 4% to 5% depending on sex. This indicates that environmental factors are more important for ferritin recovery after blood loss than for naive ferritin level. A plausible explanation for this difference is that since both exposure to air pollution and donating blood causes significant disruptions to iron homeostasis, these disruptions may interact and together have a larger effect than simply additive.

SEM is a technique well-suited to test hypotheses on how different factors interact and correlate with a specific outcome like ferritin levels, especially when there are many factors to consider. Compared to multiple (linear) regression, more complex models can be tested, and for each variable measurement error is taken into account. [58] Moreover, the percentage of variance explained by groups of related variables can be calculated and compared. The stratified approach in this study also adds to the model validity: parameter estimates can be compared across groups, allowing discovery of implausible results. Our analyses show that the convergent validity of the individual characteristics construct is lower for active donors than for new donors. This may indicate that new donors are a more homogenous group than active donors, which is likely due to the more narrow age range of new donors. Other strengths of this study are its large sample size and collection of data throughout the country.

Two main limitations of this study should be noted: its generalisability and its restricted scope. One might be tempted to generalise the results of new donors to the general Dutch population, as these donors have never donated blood before. However, even new donors form a very specific, generally healthier subgroup of the general population, which means that selection bias has likely been introduced. We can speculate that less healthy individuals would show a higher rate of inflammation, which may cause higher serum ferritin levels. On the other hand, iron deficient or anaemic individuals are likely underrepresented in our sample. As this selection bias most likely reduced variance in ferritin levels, this may have attenuated our results.

Regarding the scope, data on some other potentially important determinants of ferritin levels were not available in this study, the two most important being genetics and diet. [40, 41] Several genetic polymorphisms that have an effect on iron pathways have been identified, and these are likely to play a role in the recovery speed of ferritin levels after blood donation. [43, 59, 60, 61] Dietary behaviour, and in particular heme iron intake, is also a determinant of iron status in donors. [40, 45] Information on iron supplementation was also not available for this study. Sanquin does not prescribe oral supplementation of iron to donors, and only a small minority (8.7%) uses iron supplements. [40] Information on donors' smoking status is also expected to add value to the model. Had these determinants been available for our analysis, the proportion of variance explained in donor ferritin levels would likely have increased.

This study presents a model to explain variance in ferritin levels in individuals with or without donation history, based on three types of determinants. The model explained a relatively large part of the variance, especially in active donors. Individual characteristics and donation history form the most important determinants of ferritin levels. Although environmental factors accounted for less variance than the individual and donation history constructs, their contribution is meaningful and statistically significant. When clinicians or researchers use serum ferritin as a proxy for total body iron, they should be aware of this potentially confounding effect.

For blood services that are considering implementing ferritin testing for their donors, these results are of particular value. The results can be of use while the blood service is deciding on a sensible threshold for donation: rather than implementing a one-size-fits-all threshold, environmental conditions in the country can be taken into account. If there is a high level of air pollution, ferritin levels are likely to be overestimated, and thus a higher threshold for donation may be desired. It could even be taken further to make ferritin thresholds more tailored to a specific donor, by taking into account a donor's individual characteristics.

Appendix

Structural equation modelling (SEM) comprises a set of statistical methods that enables researchers to assess the support for hypothesised relationships between variables of interest. Its purpose is to account for variation and covariation of the variables in the model. Many different techniques are included in SEM, this appendix explains the approach taken in this particular study. In SEM, observed variables and latent constructs are distinguished. Observed variables are variables in the traditional sense, which are observations in the data set that have been collected by the researcher. Latent constructs are theoretical concepts that cannot be measured, but must be inferred from the observed variables; a well-known example is the latent construct intelligence that cannot be measured directly, but can be inferred from observed variables such as scores for an IQ test. Intuitively, observed variables that belong to a latent construct represent the same underlying concept, and latent constructs form in a way a dimensionality reduction of the observed variables. Mathematically, latent constructs represent shared variance of the observed variables related to the construct they belong to.

SEM is composed of two main model components: the measurement model, which shows how observed variables are divided among latent constructs, and the structural model, which shows the relationships between latent constructs and outcome variable(s). First, the measurement model is specified, and test its validity using confirmatory factor analysis (CFA). Often, several measurement models are tested and compared to see which division into latent constructs best fits the data. When the measurement model is considered to have a good fit, the structural part of the model is added, and the model fit is assessed for the full SEM model.

Measurement model

The validity of the latent constructs must be measured in two ways: each construct must have convergent and discriminant validity. Convergent validity occurs when the observed variables belonging to the latent construct share a high proportion of their variance. This is assessed by the factor loadings of the observed variables onto the latent construct: the higher the (absolute value of the) factor loading, the stronger the indication that this variable belongs to this construct. Very generally speaking, factor loadings greater than 0.4 are acceptable for including a variable within a construct, but this threshold depends greatly on the hypothesised interpretation of the latent variable. Variables with low factor loadings are excluded from the construct.

The discriminant validity of a latent construct is a measure for how well the construct can be distinguished from the other constructs in the model. It is measured by the covariances between latent constructs. A high covariance between two constructs can indicate that these constructs are (partly) overlapping, and thus have no discriminant validity.

If convergent and discriminant validity are satisfactory, model fit indices can be calculated for the measurement model. Commonly used indices are the chi-square test, comparative fit index (CFI), Tucker-Lewis index (TLI) and root mean square error of approximation (RMSEA). The CFI and TLI are both relative measures of fit, and compare the fit of the tested model against a null model, which in CFA means that the means and variances of each variable are freely estimated, but no correlations are included. CFI and TLI are on a scale from 0 to 1, with higher values indicating a better fit of the hypothesised model relative to the null model. The TLI is always more conservative (lower value) than the CFI, because the TLI includes a harsher penalty for the number of parameters estimated. Because the two fit indices are highly correlated, only one should be reported. We chose the TLI because of its more elegant penalty for complexity. Values higher than 0.95 indicate good fit.

The RMSEA is an absolute measure of fit that is not sensitive to large sample sizes, unlike the chi-square test. It uses the covariance matrix of the entire data set and of the fitted hypothesised model, and calculates the differences between these two. This results in a measure between 0 and 1, with lower values indicating smaller differences and better model fit. Cut-offs of 0.08, 0.05, and 0.01 indicate mediocre, good, and excellent fits, respectively.

If multiple measurement models are compared, as in this study, the best fitting model is selected, based on the fit indices described above. If these indicate sufficient model fit, the analysis can be continued with inspection of residual correlation between observed variables. If the pairwise residual correlation between two variables is high (absolute value of 0.1 or higher is a common cut-off), this indicates that these two variables share more variance than is currently captured in the model. If this occurs, the researcher needs to decide whether a covariance term for these two variables should be included in the model. However, this should only be done if there is sufficient theoretical support for an interpretable correlation between these variables. Otherwise there is a risk of overfitting the model to the data; after all, in confirmatory factor analysis we build upon a set of relationships that are hypothesised by the researcher. It is not a data-driven method of finding the best set of relationships. If such an approach is desired, exploratory factor analysis (EFA) can be applied instead of CFA.

Structural model

The structural component is added to the model once the latent constructs are defined, variables with low factor loadings are removed, and necessary covariance terms are added. The structural component consists of the relationships between latent constructs, or between latent constructs and outcome variable(s). With this, we now have three types of parameters for which an estimate must be calculated:

- Factor loadings (observed variable → latent construct);
- Covariances (observed variable ↔ observed variable);
- Regression coefficients (latent construct → latent construct or outcome variable).

Each parameter adds one degree of freedom to the model, and the number of parameters determines the identifiability of the model. Parameter estimates can only be obtained when the number of free parameters (the number of *unknowns*) is equal to or smaller than the number of independent elements in the covariance matrix of the data (the number of *knowns*), which is equal to $k(k+1)/2$, where k is the number of observed variables in the model. If there are more unknowns than knowns, the model is under-identified and no solution can be found. If the numbers are the same, the model is just identified, and a unique solution can be obtained. If there are fewer unknowns than knowns, we have an over-identified model, which means that there is no unique solution but multiple, and we can select the best solution based on fit measures. An over-identified model is desired.

In most software packages parameter estimates are obtained by a maximum likelihood estimator by default, but alternative estimators can be chosen as well. In this study most observed variables did not follow a normal distribution, which violates maximum likelihood estimator assumptions. Therefore, the diagonally weighted least squares (DWLS) method was used instead, which is more robust and provides more accurate parameter estimates in case the normality assumption is violated.

If the model is over-identified, fit measures can be reported along with the parameter estimates. Again, TLI and RMSEA are used to assess model fit, with the same thresholds as seen in the CFA (TLI > 0.9, RMSEA < 0.08). If the model fit is acceptable the parameter estimates can be interpreted. The interpretation of the parameter estimates depends on the specification of the model. By default, one factor loading in each latent construct is set to 1, to fix the scale of the latent construct. However, in order to compare factor loadings across constructs it is useful to consider standardised parameter estimates. The variance of the latent construct is then set to 1

and factor loadings are interpreted in terms of a change in variance. In this study, we look only at the standardised parameter estimates, as we are interested in the relative importance of each observed variable and latent construct.

Factor loadings indicate how much variance of an observed variable is shared with the variance of its latent construct. Higher absolute values indicate more shared variance, and the sign of the factor loading specifies the direction of the association. Covariance terms provide the same information for two observed variables, which can belong to the same construct or to different constructs. If they belong to the same construct, a high covariance term indicates that these two variables share more variance with each other than can be explained by the latent construct. Regression coefficients indicate how much variance of the outcome variable is explained by the variance of the latent construct. To find the relative effect of a single observed variable on the outcome variable, its factor loading must be multiplied by the regression coefficient that connects the latent construct to the outcome.

CHAPTER

5

Challenges and limitations in clustering blood donor hemoglobin trajectories

Published in: International workshop on advanced analysis and learning on temporal data, 72-84. doi:10.1007/978-3-030-39098-3_6

Authors: M Vinkenoog, MP Janssen, M van Leeuwen

Abstract

Background - In order to prevent iron deficiency, Sanquin measures a blood donor's hemoglobin level before each donation and only allows a donor to donate blood if it is above a certain threshold. In around 6.5% of blood bank visits by women, the donor's hemoglobin level is too low and the donor is deferred from donation. For visits by men, this occurs in 3.0% of cases. To reduce the deferral rate and keep donors healthy and motivated, we would like to identify donors that are at risk of having a low hemoglobin level. To this end we have historical hemoglobin trajectories at our disposal, i.e., time series consisting of hemoglobin measurements recorded for individual donors.

Methods - As a first step towards our long-term goal, in this paper we investigate the use of time series clustering. Unfortunately, existing methods have limitations that make them suboptimal for our data. In particular, hemoglobin trajectories are of unequal length and have measurements at irregular intervals. We therefore experiment with two different data representations. That is, we apply a direct clustering method using dynamic time warping, and a trend clustering method using model-based feature extraction. In both cases the clustering algorithm used is k-means.

Results - Both approaches result in distinct clusters that are well-balanced in size. The clusters obtained using direct clustering have a smaller mean within-cluster distance, but those obtained using the model-based features show more interesting trends. Neither approach results in ideal clusters though. We therefore conclude with an elaborate discussion on challenges and limitations that we hope to address in the near future.

Introduction

Sanquin is the national blood bank in the Netherlands. Every year, about 300 000 donors visit the blood bank, resulting in over 420 000 donations a year. Women are allowed to donate up to three times a year, men up to five times. There are many policies in place to ensure that the blood products that are collected are safe for the patients they will be given to. Moreover, Sanquin has the responsibility to prevent volunteer blood donors from developing health problems related to blood donation. One big risk of regular blood donation is anemia due to low iron stores or iron deficiency. A whole blood donation takes about 500 mL of blood from the donor, which contains 210 to 240 mg iron bound to hemoglobin. The total concentration of iron in the human body is approximately 38 mg/kg body weight for women and 50 mg/kg body weight for men, so a single blood donation constitutes a substantial loss of iron. [16, 17]

To prevent donors from becoming iron deficient, their hemoglobin levels are checked before each blood donation. Based on the hemoglobin measurement it is decided whether they may donate at that time: the lower limit for donation is 7.8 mmol/L for women, and 8.4 mmol/L for men. When a donor is below the threshold, they are sent home and can return for donation a few weeks later. This type of deferral occurs quite frequently: about 6.5% of female and 3.0% of male donors have too low hemoglobin levels when they visit the blood bank.

The large number of deferrals is problematic, both for donors and the blood bank: being deferred from donation is demotivating for the donor, who may decide not to return in the future, and not efficient for the blood bank, leading to a higher cost per blood product.

Because of this, Sanquin and other blood banks internationally spend considerable resources on investigating ways to reduce the deferral rate while keeping donors healthy. One asset that can be exploited for this are the hemoglobin measurements that blood banks have recorded in the past. In this paper we report on a preliminary study investigating whether we can distinguish groups of donors having different trends in their hemoglobin trajectories; if this is the case, these trends could be used to devise more personalised invitation and deferral policies.

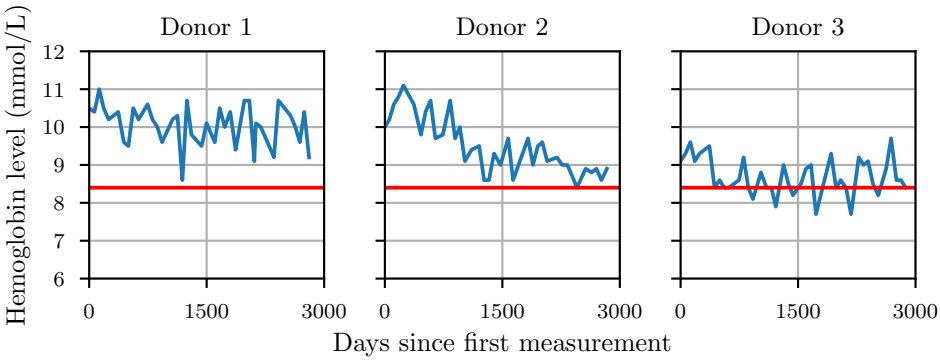

Figure 5.1: Hemoglobin trajectories of three male donors. From left to right: a high stable trajectory, a declining trajectory, and a low stable trajectory. The red line is the hemoglobin threshold for donation (8.4 mmol/L for male donors).

5

Approach and contributions

We have data available on all blood bank visits in the Netherlands since 2006. For every donor, we have only two relevant background variables: year of birth and sex. It has long been known that age and sex affect hemoglobin levels. Men's levels are higher than women's and decrease with age, while women's levels increase after menopause. [62]

Apart from these factors, a large part of the variation in hemoglobin levels can be attributed to diet and lifestyle: the iron richness of the donor's diet and their activity level play a substantial role here. However, we don't have large-scale data on this. The clusters of donors we hope to identify could be a proxy for these variables.

The more interesting part of the data are the hemoglobin measurements taken every time the donor visits the blood bank. Each measurement has a time stamp, and together the individual measurements of a single donor form a time series; we will refer to these time series as hemoglobin trajectories.

We aim to find groups of donors whose hemoglobin levels are similar throughout their donation history. More specifically, we would like to distinguish donors with a stable (high or low) hemoglobin level from donors with a declining level over time, as these require different donation policies. The three different trends that we expect to find in the data are illustrated in Figure 5.1.

Finding groups of similar data points in an unsupervised manner is a typical clustering task and since hemoglobin trajectories are time series, we naturally resort to

time series clustering. Time series clustering can be applied in many fields and has been studied for a long time, as a result of which a large number of clustering methods for time series exist. [63, 64, 65]

A big limitation, however, is that most existing algorithms require the time series to be sampled at fixed, equidistant time stamps. In our data, the sampling intervals are highly irregular on two levels. First, the intervals are not uniform across time series; an easy example is that women are allowed to donate three times a year, men five times. Second, the intervals are not uniform within the time series either: sometimes a donor returns for their next donation two months after the previous one, sometimes six months. Donors can also temporarily stop donating, and then return years later. A related limitation that is relevant to our data is that the time series have unequal lengths. Many donors in the data set have been regularly donating for over ten years, while others have just started.

Faced with these challenges, in this paper we will investigate whether we can transform our data for use with a standard clustering method without losing critical information. Specifically, we will employ two approaches:

1. Direct clustering using re-sampling combined with dynamic time warping [66] as distance measure;
2. Trend clustering using model-based feature extraction combined with the Euclidean distance.

As our main aim is to evaluate and compare the data representations, the choice of a clustering method is less important; we will use k-means because it is straightforward, effective, and well-known. [67]

The main contributions of our preliminary study are a proof-of-concept showing that clustering of hemoglobin trajectories of Dutch blood donors is feasible, and the identification of challenges and limitations of using time series clustering for hemoglobin trajectories. We consider these to be important first steps towards an effective clustering method for irregular time series in which the irregularities itself may contain useful information.

Data

Our data consists of all blood donations made at any of Sanquin's locations between January 2006 and June 2018, extracted from the blood bank's database system eProgesa. In total, there are 6 945 611 donations by 688 665 unique donors. Because we

are interested in donors' hemoglobin trajectories from their first donation onward, we selected for our analyses all donors that did not visit the blood bank before 2010. It is possible that there are donors in the data set that donated before 2006 and returned after a gap of at least four years, but we expect this number to be low, and their hemoglobin levels similar to actual new donors.

Many types of blood donation take place at Sanquin, the most common being plasma donation and whole blood donation. During plasma donation, red cells are returned to the donor and only the plasma is collected. As hemoglobin is contained in the red blood cells, this type of donation does not have a substantial effect on hemoglobin levels. Therefore, we only look at donors that donate whole blood, during which no blood components are returned to the donor.

We take into account donors that have donated whole blood at least eight times in our time window—once a year on average. There are 23 856 female and 20 299 male donors that fit these criteria. To decrease computation time, we randomly selected 5000 women and 5000 men for our experiments. Within this data set, the deferral rate due to low hemoglobin is 7.8% among female donors and 3.3% among male donors.

The two data sets contain 5000 individual univariate time series each, consisting of the hemoglobin measurements during the visits to the blood bank. Hemoglobin is measured in mmol/L. The median number of measurements per time series is 12 for women (interquartile range, IQR 10–14) and 14 for men (IQR 11–19).

The time intervals between measurements differ both within and between time series. The minimum required interval between two donations is 122 days for women and 56 days for men, but it can even be a few years. The median interval for women is 133 days (IQR 112–169) and for men 79 days (IQR 64–114). Aside from the hemoglobin measurements, the only variable used is the sex of the donor. Clustering methods will be applied separately to the female and male subsets.

Methods

We will experiment with two data representations and compare the results of the k-means clustering algorithm on both representations. The methods will be compared on cluster tightness using mean within-cluster distance, and visually on the informativeness of the cluster using the graphs of the cluster centroids.

The first method employs direct clustering using dynamic time warping based on the hemoglobin levels at each time point, the second method employs trend clustering using model-based feature extraction. Preprocessing is the same for both.

Preprocessing

When time series are of equal length and have the same measurement intervals, clustering is relatively straightforward. At each time point, we can calculate the difference between measurements in two time series, and group time series with smaller differences in the same cluster. However, from this perspective our data is rather messy: time series are all of differing lengths and have different measurement intervals, both within and between individuals. While there are more sophisticated ways to handle this (see the Discussion), none of the existing algorithms that we found are perfectly suited to our data. Therefore, for this first trial we decided to side-step the problem of unequal intervals by resampling the time series to regular intervals by linear interpolation.

We take each donor's first donation since 1 January 2010 as the starting point of their time series. All time stamps are relative to the first time stamp, recorded as days since first donation. Hemoglobin values are then resampled to weekly measurements using linear interpolation. This gives a maximum of 439 observations per donor, one for each week between 1 January 2010 and 1 June 2018. Donors that started donating later in the time window will have fewer measurements, and thus have a number of missing values at the tail of the time series. For the first 140 weeks, the number of donors with missing values is almost zero, but then the number of donors that still has measurements starts dropping at a steady rate. We chose to use hemoglobin measurements up to 286 weeks after the first donation, at which time half of our 5000 donors has no missing values, and the other half misses at most 50% of observations.

Direct clustering using dynamic time warping

For this method, the features that we will feed to the clustering algorithm are the resampled hemoglobin measurements as described in the previous section. As a distance measure, we use dynamic time warping (DTW) with the window parameter set to $w = 5$. [66] This algorithm is better-suited to our data than for instance the Euclidean distance, because it takes into account varying speeds and time shifts. Because the time series vary in length, we compare time series only up to the last data point in the shortest series.

The algorithm can be summarised as follows:

1. Calculate the Euclidean distance between the first point in the first series, and every point within the window of $w = 5$ in the second series;
2. Store the minimum distance calculated;
3. Repeat steps 1–2 for all points in the first series;
4. Add all the minimum distances to get the DTW distance.

Trend clustering using model-based feature extraction

The second method takes as input for the clustering algorithm not the (resampled) time series itself, but rather a set of features that should summarise the time series in such a way that similar time series will have similar feature values. We are interested in distinguishing three types of hemoglobin trajectories: high stable, low stable, and declining. We therefore choose to cluster the trajectories based on the intercept and slope of the linear trend.

The intercept and slope are calculated using linear least-squares regression on the resampled time series described in the previous section, to allow for a direct comparison between the two methods. Because the slope and intercept values are on different scales, we normalise them using a min-max scaler before clustering. The values are then all between 0 and 1, 0 being the minimum value among the time series and 1 the maximum.

Clustering algorithm

For the actual clustering, we use k-means clustering, a heuristic algorithm that is usually quite fast at finding a local optimum. [67] It requires the user to specify the number of desired clusters k. We chose this well-known algorithm for its wide applicability and straightforward implementation.

For the direct clustering, the input to the algorithm contains the resampled time series. Because of the differing lengths of the time series, we chose to initialise the clusters randomly from a uniform distribution, instead of choosing k time series as initial cluster centroids. The distance measure used is DTW.

For the trend clustering, the input consists of two features per trajectory: the intercept and the slope of the linear trend. As distance measure the Euclidean distance is used.

In general, k-means clustering returns the best results if the algorithm stops when the difference between the cluster centroids in two subsequent iterations is smaller than some ϵ. Because the program is computationally expensive due to the DTW calculations, we opted to let it run for at most five iterations for the first clustering method.

The algorithm is as follows:

1. Initialise k cluster centroids;
2. Assign each time series to the cluster to which it is most similar, based on the specified distance;
3. Recalculate the cluster centroids by taking the average value for each feature;
4. Repeat steps 2–3 for 5 iterations or until convergence.

Evaluation

We compare the clusters based on the two data representations in two ways: cluster tightness and cluster informativeness. The first is a numerical comparison, the second graphical. Cluster tightness is assessed by the mean within-cluster distance. For each cluster, we calculate the distance from the cluster centroid to the individual time series by taking the DTW distance between the two. The mean of these distances is the mean within-cluster distance. We also calculate the sum of the within-cluster distance for each value of k, which is the sum of the DTW distances between the individual time series and the cluster centroids, summed over all clusters. As the number of clusters increases, we expect the sum of the within-cluster distances to decrease.

Cluster informativeness is assessed visually by looking at the graphs of the cluster centroids. We hope to see centroids that are different in slope, and not just horizontal lines with different average hemoglobin values.

Results

We will first present the results from both methods separately, then compare the two on cluster tightness and informativeness.

Direct clustering

Figure 5.2 shows the centroids of the clusters after direct clustering with DTW. At $k = 2$ and $k = 3$, we see that the clusters are based mostly on the mean hemoglobin

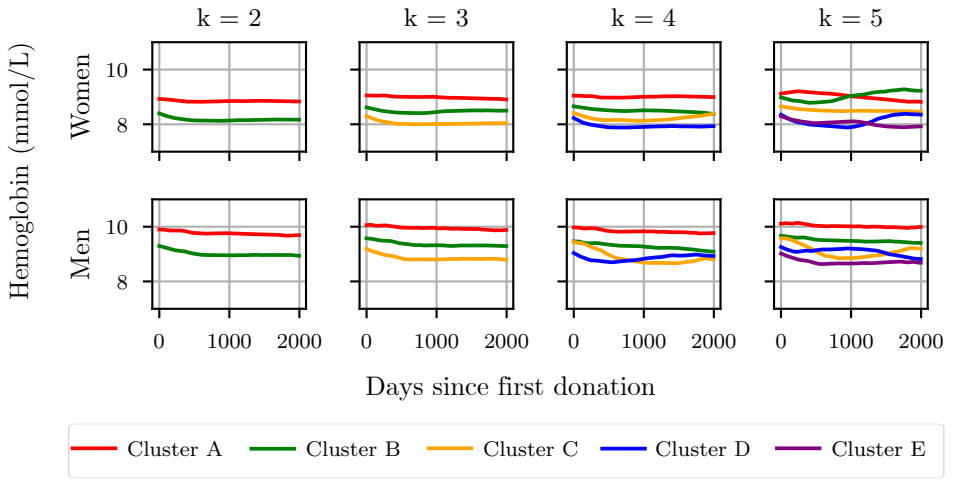

Figure 5.2: Cluster centroids after clustering resampled hemoglobin trajectories of 5000 female and 5000 male donors with the k-means clustering algorithm ($k \in [2, 3, 4, 5]$) and DTW distance as distance measure.

level in the donors, and cluster centroids are almost parallel. At higher numbers of clusters, we start to see some differences in trends as well, with centroids intersecting each other. At $k > 5$, we saw that centroids start overlapping for longer periods of time and are no longer distinct enough to be informative. These graphs are not included in the paper. In almost all centroids, there is a decrease in hemoglobin value at the beginning of the hemoglobin trajectory.

To assess the tightness of the clusters, Table 5.1 shows the mean within-cluster distances, with DTW used as distance measure. The total sum of the within-cluster distances decreases as the number of clusters increases, which is expected because the same distance measure was used to create the clusters. The names of the clusters correspond to those in Figure 5.2. Table 5.1 also shows the number of time series assigned to each cluster. We see that in size, the clusters are quite well-balanced: the smallest cluster has size 413 where size 1000 would be expected if all clusters were the same size (female donors, $k = 5$, cluster B).

Trend clustering

Figure 5.3 shows the cluster centroids after clustering on trend features. As after the direct clustering, the centroids are distinct from each other and do not intersect at $k = 2$ and $k = 3$. From $k = 4$ and up, cluster B shows an interesting new trend in

Sex	k	Cluster A $\bar{d}$ (N)	Cluster B $\bar{d}$ (N)	Cluster C $\bar{d}$ (N)	Cluster D $\bar{d}$ (N)	Cluster E $\bar{d}$ (N)	Sum $\bar{d}$
Women	2	7.1 (1613)	6.3 (3387)				32670
	3	5.7 (2128)	7.0 (986)	5.9 (1886)			30135
	4	5.4 (1197)	5.2 (1205)	5.9 (1671)	6.9 (927)		29049
	5	6.3 (475)	6.5 (413)	5.5 (1379)	5.9 (1766)	5.4 (967)	28840
Men	2	7.2 (2020)	6.3 (2980)				33260
	3	6.1 (1997)	5.6 (1851)	6.9 (1152)			30508
	4	7.1 (1600)	5.8 (1589)	5.5 (835)	5.1 (976)		30222
	5	4.9 (896)	5.9 (1424)	5.2 (906)	5.5 (871)	6.8 (903)	28567

Table 5.1: The mean distance from the centroid to the time series ($\bar{d}$) and the number of time series in each cluster (N) after direct clustering. Dynamic time warping is used as distance measure. The rightmost column shows the sum of the within-cluster distances.

male donors: the slope of the line is much steeper than those of the other clusters.

In Table 5.2, we see that the mean distance from the centroid to the individual time series is larger than in the clusters obtained using the first method. The sum of the within-cluster distances does not decrease as k increases, and for female donors it even increases substantially. This can happen because in this method, the clusters are decided based on the Euclidean distances between the trend features of the time series, rather than the DTW distance between time series as in the first method.

The number of time series per cluster is mostly well-balanced, although there are some cases of small clusters: at $k = 5$, in male donors, cluster A only contains 386 time series where 1000 would be expected if all clusters were of equal size.

Comparison

From the within-cluster distances, it is clear that the direct clustering method leads to tighter clusters. Figure 5.4 illustrates this well. It shows the result of both direct and trend feature clustering on male donors with $k = 4$ clusters. Each subplot shows the cluster centroid in red, and 100 randomly selected individual time series within the cluster in grey. Although after both direct and trend clustering the cluster centroid lies in the middle of the individual time series, the spread is much smaller in direct than in trend clustering.

In both methods, cluster centroids vary mostly in the average hemoglobin value over time, and not as much in trend, which is what we are mostly interested in. The exception is cluster B in the trend clustering method, which shows a relatively steep

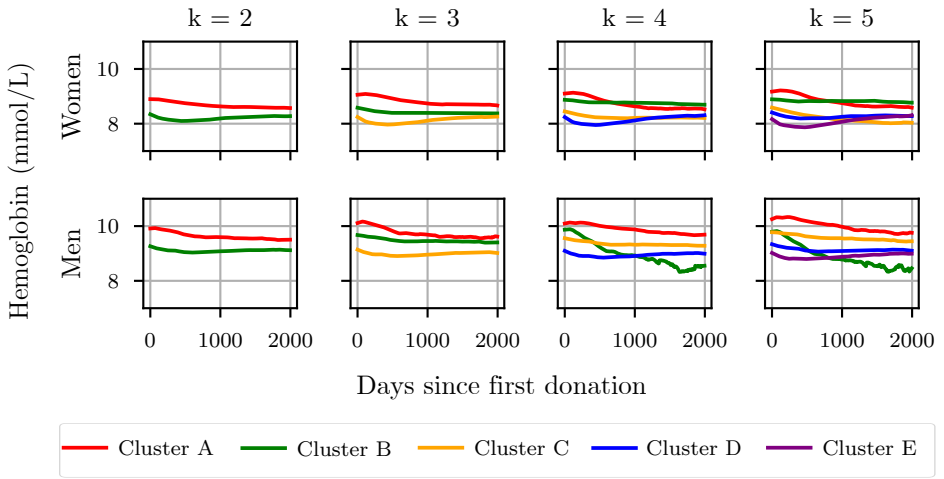

Figure 5.3: Cluster centroids after clustering resampled hemoglobin trajectories of 5000 female and 5000 male donors based on the intercept and slope of the linear trend, using the k-means clustering algorithm ($k \in [2, 3, 4, 5]$).

Sex	k	Cluster A $\bar{d}$ (N)	Cluster B $\bar{d}$ (N)	Cluster C $\bar{d}$ (N)	Cluster D $\bar{d}$ (N)	Cluster E $\bar{d}$ (N)	Sum $\bar{d}$
Women	2	8.1 (2028)	6.8 (2972)				36689
	3	10.1 (1075)	6.5 (1727)	9.6 (2198)			43195
	4	8.2 (602)	10.8 (1181)	6.7 (1362)	10.5 (1855)		46318
	5	6.2 (839)	14.1 (1016)	8.5 (881)	11.9 (1761)	8.8 (503)	52431
Men	2	11.0 (2843)	11.1 (2157)				55156
	3	9.5 (831)	6.5 (1924)	8.7 (2245)			40113
	4	10.8 (389)	15.3 (961)	7.0 (2104)	6.4 (1546)		43392
	5	6.0 (386)	6.3 (1071)	7.1 (1378)	7.9 (445)	8.6 (1720)	37234

Table 5.2: The mean distance from the centroid to the time series ($\bar{d}$) and the number of time series in each cluster (N) after trend clustering. Dynamic time warping is used as distance measure for evaluation. The rightmost column shows the sum of the within-cluster distances.

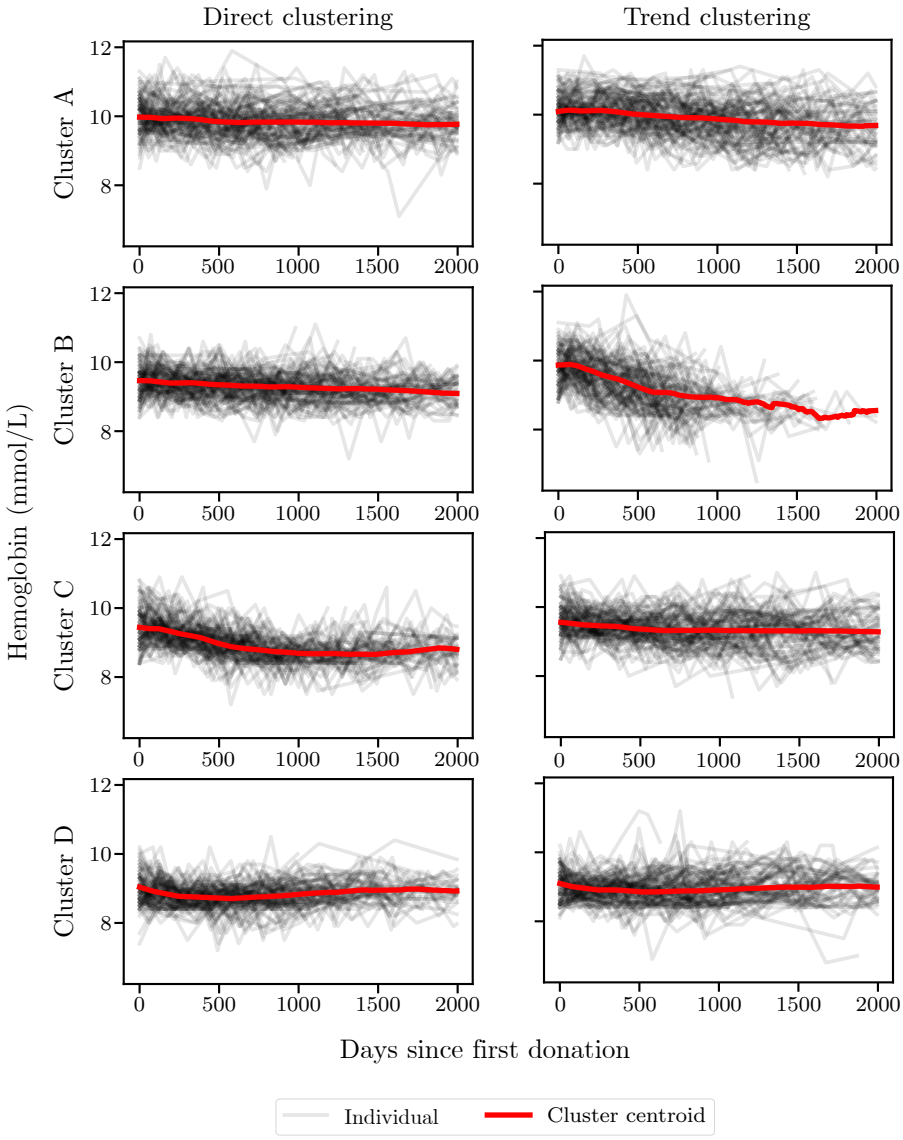

Figure 5.4: Cluster centroids after clustering resampled hemoglobin trajectories of 5000 male donors, using direct (left) or trend clustering (right) and k-means clustering with $k = 4$. Red lines are the cluster centroids. 100 randomly sampled individual hemoglobin trajectories from each cluster are also plotted to show the fit.

downward trend.

To verify the stability of the cluster centroids obtained by the k-means algorithm, we ran it several times with different random initialisation values. Visual inspection of the results showed that the algorithm consistently converged to the same centroids.

Discussion

The clusters obtained by the two methods are clearly very different. The centroids of the clusters are much more linear when the direct clustering is applied, compared to the trend feature clustering. The mean within-cluster distances are much smaller in the first method, which indicates denser clusters. However, this comparison is biased, because the first method used the same distance measure during clustering, so it is expected to minimise this distance. The second method minimised the Euclidean distance between the linear trend features of the time series, and not the DTW distances.

The results from the direct clustering are in line with our expectations. The clusters suit the time series relatively well, and the total sum of within-cluster distances decreases as the number of clusters increases. However, the time series are clustered together mostly on average hemoglobin level, which is not what we are looking for in this context. We would prefer to identify clusters based on the overall trends in each donor, so that we can distinguish donors with a high stable hemoglobin level, a low stable hemoglobin level, and decreasing hemoglobin levels from each other.

This is what we expected to see after clustering time series based on trend. It is partly what we see in the cluster centroids: in male donors, for $k = 4$ and $k = 5$, cluster B is very distinct from the others and has a steep downward slope. We know that declining hemoglobin trajectories are highly prevalent in female donors as well, but none of those centroids have a slope close to the one in male donors.

An interesting observation is that in almost all clusters, the hemoglobin level is decreasing in the first ± 500 days and then plateaus. This indicates that there is an initial effect of blood donation on average hemoglobin levels, but after the initial effect it reaches a new steady state. However, this is only based on the average hemoglobin levels of 5000 donors, and individual trajectories still show a lot of variation over time, making it hard to predict.

Limitations

There are some features in the data that were ignored in this first exploration in hemoglobin trajectory clustering. There is a seasonal component to hemoglobin levels: in warm seasons, levels are lower than in cold seasons. Because we used the number of days since first donation as time points and not the actual dates, we lost this information. An improvement would be to correct for seasonal variations before transforming the time variable. The same applies for the time of day hemoglobin was measured: it is highest in the morning and then drops steadily throughout the day.

A very clear feature of the data that was not used is the unequal sampling interval. Both methods required the intervals to be equal, so we resampled the time series using linear interpolation to satisfy this requirement. This means that we lose the information contained within the sampling intervals, and the resampled data points are of lower accuracy than the original measurements.

The third feature of the data that we would like to include in further analyses is whether or not a donation followed the hemoglobin measurement. If the hemoglobin level is below the threshold of 7.8 mmol/L for women or 8.4 mmol/L for men, no donation is made, and it is likely that the next measurement is higher. There is also an interaction with the interval length: if a donor has donated blood, the next measurement has to be at least 56 days later, but if the hemoglobin level was too low, it can be shorter.

Other Irregular Time Series Frameworks

There are many more fields in which irregular time series are observed (astronomy, medicine, economics, etc.), and in which the irregularities contain information we don't want to lose by transforming the data to equally spaced data. Some algorithms focus on calculating rolling time series operators such as simple moving averages or exponential moving averages. [68] This is a more elegant form of interpolation than what we have applied here, but the information contained in the intervals themselves is still lost.

A more fitting approach for our data might be a framework that takes two time series as input for each donor: one containing the hemoglobin measurements and one containing the interval lengths. We might consider a move to more complex algorithms, such as recurrent neural networks (RNNs) in combination with long short-term memory (LSTM) cells. [69] While the majority of RNN implementations still uses fixed time steps, the Phased LSTM model, which introduces an additional time

gate, handles irregular intervals without losing the information contained within the time steps. [70] A similar approach is Time-LSTM, which has been used to model website users' sequential actions by taking into account the sampling intervals. [71]

Another deep learning model that looks at informative missingness is GRU-D [72], which is based on gated recurrent units (GRU). It has been applied to real-world clinical data sets, where the missingness rate is highly correlated with variables of interest. This model has achieved good results in supervised classification tasks, and may also have useful applications for our unsupervised clustering task.

Future Work

By clustering donors' hemoglobin trajectories we hope to find clusters of donors that respond similarly to frequent blood donation. We assume that the clusters are a proxy for unobserved donor characteristics, such as iron intake, diet, physical activity levels and iron needs. If clustering is successful, we want to search for correlations between the cluster and donor information collected in questionnaires in previous studies carried out at Sanquin (Donor InSight). Eventually, the goal is to predict as early as possible in a donor's donation career which cluster they belong to, and to assign an optimal donation frequency based on this information. That way, deferral due to low hemoglobin may be minimised, and donors will stay healthy and motivated.

CHAPTER

6

Associations between symptoms, donor characteristics and IgG antibody response in 2082 COVID-19 convalescent plasma donors

Published in: Frontiers in Immunology 13: 821721. doi:10.3389/fimmu.2022.821721

Authors: M Vinkenoog, M Steenhuis, A ten Brinke, JG van Hasselt, MP Janssen, M van Leeuwen, FH Swaneveld, H Vrielink, L van de Watering, F Quee, K van den Hurk, T Rispens, B Hogema, CE van der Schoot

Abstract

Background - Many studies already reported on the association between patient characteristics on the severity of COVID-19 disease outcome, but the relation with SARS-CoV-2 antibody levels is less clear.

Methods - To investigate this in more detail, we performed a retrospective observational study in which we used the IgG antibody response from 11 118 longitudinal antibody measurements of 2082 unique COVID convalescent plasma donors. COVID-19 symptoms and donor characteristics were obtained by a questionnaire. Antibody responses were modelled using a linear mixed-effects model.

Results - Our study confirms that the SARS-CoV-2 antibody response is associated with patient characteristics like body mass index and age. Antibody decay was faster in male than in female donors (average half-life of 62 versus 72 days). Most interestingly, we also found that three symptoms (headache, anosmia, nasal cold) were associated with lower peak IgG, while six other symptoms (dry cough, fatigue, diarrhoea, fever, dyspnoea, muscle weakness) were associated with higher IgG concentrations.

Introduction

Severe acute respiratory syndrome corona virus 2 (SARS-CoV-2) emerged late 2019 in China, and by March 2020 was declared a pandemic by the World Health Organization (WHO). As of September 2021, over 200 million individuals have been infected with COVID-19, which has inflicted an immense impact on the healthcare system worldwide. The virus mainly targets the respiratory tract, which can lead from mild symptoms to severe respiratory distress syndrome. Studies have shown that antibody responses against the SARS-CoV-2 spike protein can be first detected 1-3 weeks post symptom onset in most COVID-19 patients, [73, 74] and remain in circulation for up to 1 year. [75, 76, 77, 78] There is however a substantial variation in antibody levels between individuals. [77]

Many studies have reported on the association between disease severity and donor characteristics, such as sex, body mass index (BMI), age, and blood group. Males tend to be more susceptible to develop a severe course of the SARS-CoV-2 virus infection. [79, 80] In addition, age above 50 and obesity are also associated with increased risk of severe outcome. [80, 81, 82, 83] ABO blood type may also play a role in COVID-19 infection, but the exact influence remains unclear. [84, 85]

Antibody responses also seem to be associated with symptoms and clinical information. In general, SARS-CoV-2 antibody levels are higher in patients with a severe disease outcome. [86] A recent study in which COVID-19 convalescent plasma (CCP) donors were followed for three months after symptom resolution showed that greater disease severity, older age, male sex, and high BMI correlate with high SARS-CoV-2 antibody levels. [79, 87] The same study also reported that particularly the symptoms fever, body aches, and low appetite correlate with high SARS-CoV-2 antibody levels. Limitations of this study include a small number of subjects and the low number of longitudinal data points available for each subject, which restricts the possibilities to analyse trends in antibody levels over time and the association with donor characteristics and symptoms.

Here, we aimed to gain a more detailed insight into individual symptoms and donor characteristics and their association with the IgG antibody response. Therefore, we analysed a longitudinal data set of 11 118 anti-RBD antibody measurements of 2082 unique CCP donors. Interestingly, we found that three symptoms (headache, anosmia, nasal cold) were associated with lower peak IgG, while six other symptoms (dry cough, fatigue, diarrhoea, fever, dyspnoea, muscle weakness) were associated with higher IgG concentrations.

Methods

Study population samples

Between April 2020 and March 2021, Sanquin Blood Bank (Amsterdam, the Netherlands) collected samples from over 24 000 COVID-19 recovered adults who enrolled in the CCP programme. Within this programme, plasma is derived from patients that recovered from COVID-19, with the aim to help patients recover from COVID-19. Donation was voluntary and non-remunerated, and donors provided written informed consent before their first donation. Donors were included based on either a positive PCR or presence of anti-RBD IgG antibodies above 80 Arbitrary Units per ml (AU/ml) and after being free of symptoms for at least two weeks. Donors donated plasma on average every two weeks, until antibody levels were below 4 AU/ml in two consecutive donations. Only donors with at least three consecutive antibody measurements and a complete questionnaire were included in the analyses, resulting in a study population of 2082 donors. Supplementary Figure S6.1 shows the number of donors that were excluded at each step.

Questionnaire

Starting August 2020, donors that enrolled in the convalescent plasma programme were invited by e-mail to fill out an online questionnaire, programmed in Qualtrics (SAP, Walldorf, Germany). The questionnaire included questions about the possible origin of the infection, the reason why donors were tested and a list of 18 symptoms considered to be COVID-19-related according guidelines specified by the Dutch National Institute for Public Health and the Environment. [88] Participants could indicate if they experienced symptoms and, if the symptoms were present, how severe these symptoms were on a 4-point scale, from 1 (very mild) to 4 (severe). Additionally, participants were asked about the duration of their symptoms, whether they consulted a physician or were admitted to hospital and/or intensive care units. The full questionnaire is included as an online supplement. Donors were excluded from analysis if sex, age and/or date of illness was absent.

Antibody measurements

IgG to RBD was measured essentially as described before. [77, 78] In brief, plates were coated with recombinant RBD, incubated with samples, and bound IgG antibodies

were detected using an anti-human IgG antibody (MH16, Sanquin); quantification was done relative to a plasma pool consisting of CCP donors and expressed as AU/mL.

Statistical model

Longitudinal trends in antibody levels were analysed with a linear mixed-effects model, using log-transformed anti-RBD IgG levels as outcome variable. Timepoint 0 corresponds to 20 days post onset of symptoms. [73] As such, the estimated intercept of the model corresponds to a donor's estimated peak IgG level. [89] The estimated slope of the model is used to calculate a donor's IgG half-life, in days: $t_{1/2} = log(\frac{1}{2})/slope$.

Only measurements within six months post onset of symptoms were included, as in later stages of recovery antibody decline is expected to slow down and no longer expected to follow a loglinear decline. [77]

A three-step approach was used to analyse the effects of the covariates. In the first step, a null-model was fit to the data, using time as the only predictor variable and allowing a random intercept and slope to be estimated for each donor. In the second step, we tried to explain the variance in random intercepts and slopes by including fixed effects for donor characteristics, i.e., sex, age, height, weight, BMI, and blood group (ABO and RhD), in addition to the random intercept and slope per donor. In the third step, fixed effects that were statistically insignificant in the second step were removed and additional donor information variables obtained from the questionnaires were added as fixed effects. This information concerned data on hospitalisation, ICU admission, co-morbidities, and the presence of 18 symptoms as shown in Table 6.1. This approach allowed separate estimation of the proportion of variance explained by donor characteristics and clinical information.

Significance levels of individual variables were estimated using Satterthwaite's approximation, as degrees of freedom cannot be calculated exactly in models that include both random and fixed effects. [90] Because this approximation is slightly anti-conservative, an alpha-level of 0.01 was chosen to determine statistical significance. Non-significant predictors were excluded after each step. Relative quality, taking into account both goodness of fit and model complexity, of the models was assessed by comparing the Akaike information criteria (AIC) after each step.

Data were processed and analysed with the R programming language and environment for statistical computing (version 4.0.3), using packages lme4 and lmerTest for analyses and ggplot2 for generating graphs.

Results

Study population characteristics

We used 11 118 antibody measurements of 2082 unique donors to study the associations between symptoms, donor characteristics, and IgG antibody response. The number of available antibody measurements per donor ranged from 3 to 18 measurements. In addition, each donor completed a questionnaire, which gave insight into symptoms and donor characteristics. Table 6.1 shows the distributions of donor and COVID-19 related disease characteristics in the study population.

Compared to all active whole-blood and plasma donors in 2020, donors in our study population are slightly older (46 vs 42 years for women, 52 vs 48 years for men). Median weight and height, as well as proportion of female donors and rhesus D blood group are similar to those of the active donor population. Blood group A is overrepresented in our study population (47% vs 39% for women, 45% vs 39% for men), while blood group O is underrepresented (39% vs 47% for women, 42% vs 47% for men).

Table 6.1: Study population characteristics. Continuous variables are represented by their median and interquartile range (IQR), categorical variables by absolute count and percentage.

	Women		Men	
	Median or count	IQR or percentage	Median or count	IQR or percentage
Number of donors (proportion of total)	1236	59.4%	846	40.6%
Number of donations per donor	6	4 – 8	6	4 – 10
Days POS at first donation	48	33 – 77	47	32 – 77
Days POS at last donation*	122	97 – 151	126	103 – 157
Age (years)	45.9	28.0 – 55.3	51.8	39.6 – 59.3
Height (cm)	171	167 – 176	184	180 – 189
Weight (kg)	73	65 – 83	88	80 – 97
BMI (kg/m2)	24.8	22.6 – 28.4	26.4	24.0 – 28.2

Blood group ABO				
A	581	47%	381	45%
B	120	9.7%	84	9.9%
O	484	39%	352	42%
AB	51	4.1%	29	3.4%
Blood group RhD				
Positive	1024	83%	691	82%
Negative	212	17%	155	18%
Hospital admission	19	1.5%	50	5.9%
Intensive care	4	0.3%	8	0.9%
Symptoms				
Asymptomatic	*8*	*0.6%*	*7*	*0.8%*
Fatigue	979	79%	597	71%
Anosmia/ageusia	853	69%	471	56%
Headache	820	66%	467	55%
Myalgia	705	57%	445	53%
Nasal cold	692	56%	424	50%
Fever	621	50%	507	60%
Dry cough	560	45%	396	47%
Sore throat	519	42%	307	36%
Chills	499	40%	356	42%
Sneezing	461	37%	381	45%
Dyspnoea	461	37%	297	35%
Muscle weakness	426	34%	260	31%
Diarrhoea	221	18%	102	12%
Nausea	184	15%	72	8.5%
Sputum production	178	14%	152	18%
Altered mental status	127	10%	80	9.5%
Skin rash	69	5.6%	27	3.2%
Vomiting	49	4.0%	28	3.3%

Null-model fit (step 1)

In the first step we estimated an intercept and slope for each individual donor using the null model, describing the linear relationship between log-transformed IgG levels and time post onset of symptoms. [73] The residuals, i.e., the difference between measured IgG and predicted IgG as estimated by the null model, follow a normal distribution with mean 0 and standard deviation of 0.21 log (AU/ml). This distribution is independent of time post onset symptoms, supporting the assumption that the relationship is linear after log-transformation. Given this assumption, the estimated peak IgG level set at 20 days POS is most likely an accurate extrapolation and allows for comparisons between donors. Supplementary Figure S6.2A shows the fitted line and actual measurements for four randomly selected donors (donors A to D). Supplementary Figure S6.2B shows the distribution of the residuals over all observations for all donors.

After analysing all samples, we found a median peak IgG concentration of 38.8 AU/ml (IQR 20.9-78.6) and a median half-life of 66 days (IQR 50-94) (Figure 6.1). For the majority of donors, the estimated slope corresponds to a plausible antibody half-life. However, for 80 donors (3.8%), the fitted slope was positive, which results in a negative estimated half-life estimate. For an additional 59 donors (2.8%), the estimated half-life is extremely long (defined here as more than 365 days, but estimates ranged up to 16 000 days). This occurs when the estimated slope is very close to zero (but still negative), which may happen when IgG levels barely decrease between measurements and no decay in antibody levels are measured. Examples of donors with a negative half-life and very long half-life are given in Supplementary Figure S6.3. These donors were not excluded from the study in order not to overstate accuracy, and because there was no reason to assume the IgG measurements were incorrect.

Associations with predictor variables

The results of step 2, where individual donor characteristics were added to the model as predictor variables, are shown in Figures 6.2A–C and Table 6.2. Sex was associated with the slope (Figure 6.2A), as the rate of antibody decay is faster in men: the median slope for men corresponds to a half-life of 62 days, while this is 72 days for women. Men displayed higher peak IgG levels than women, but this difference was not statistically significant ($p = 0.68$). Age (Figure 6.2B) and BMI (Figure 6.2C) were both positively correlated with peak IgG concentration. A one-year increase in age corresponds to a 0.013 increase in the log-transformed IgG level, an increase of one BMI point corresponds to a 0.024 increase in log-transformed IgG level. No

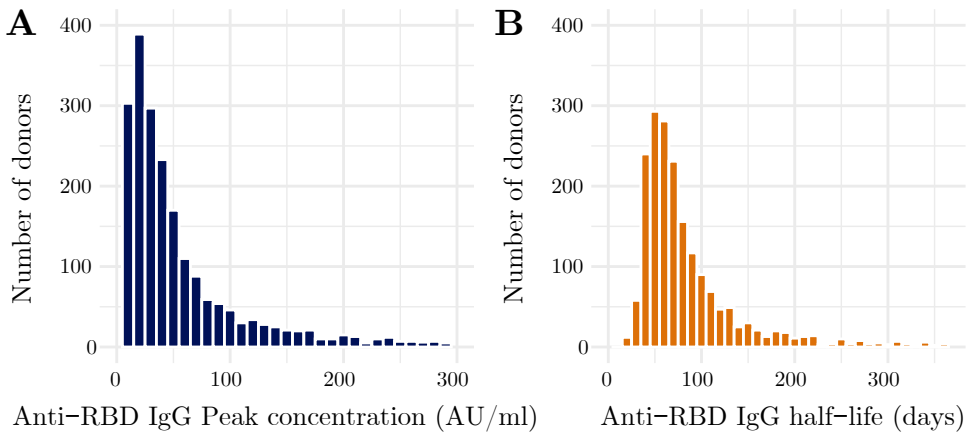

Figure 6.1: Anti-RBD IgG peak and half-life. (A) Distribution of estimated peak IgG concentration (at 20 days POS) and (B) estimated half-life of 2082 COVID convalescent plasma donors, as estimated by the null model. Please note that since both distributions have an extremely long right tail, the horizontal axes are truncated at (A) 300 AU/ml and (B) 365 days, excluding 70 and 139 donors from left and right histograms, respectively.

significant associations with antibody titres were found for variables blood group, height, and weight. Random effects for peak IgG level and antibody half-life are positively correlated with a correlation coefficient of 0.29, indicating that higher peak IgG is moderately associated with higher (less negative) slope, and therefore with a longer half-life.

Associations with clinical information

After adding clinical information significant associations with peak IgG concentration were found for hospital admission and various clinical symptoms (Figures 6.2D, 6.3, 6.4 and Table 6.2). Hospital admission was significantly associated with both higher peak IgG level and shorter half-life (Figure 6.2D). Nasal cold, headache, and anosmia were associated with lower peak IgG levels, while dry cough, fatigue, fever, dyspnoea, diarrhoea, and muscle weakness were associated with higher peak IgG levels. Figure 6.3 shows the estimated peak IgG level when these symptoms are present. Note that values on the y-axis are the predicted peak IgG levels when all continuous variables are equal to their average value, and all binary variables (hospital admission and all other symptoms) equal zero.

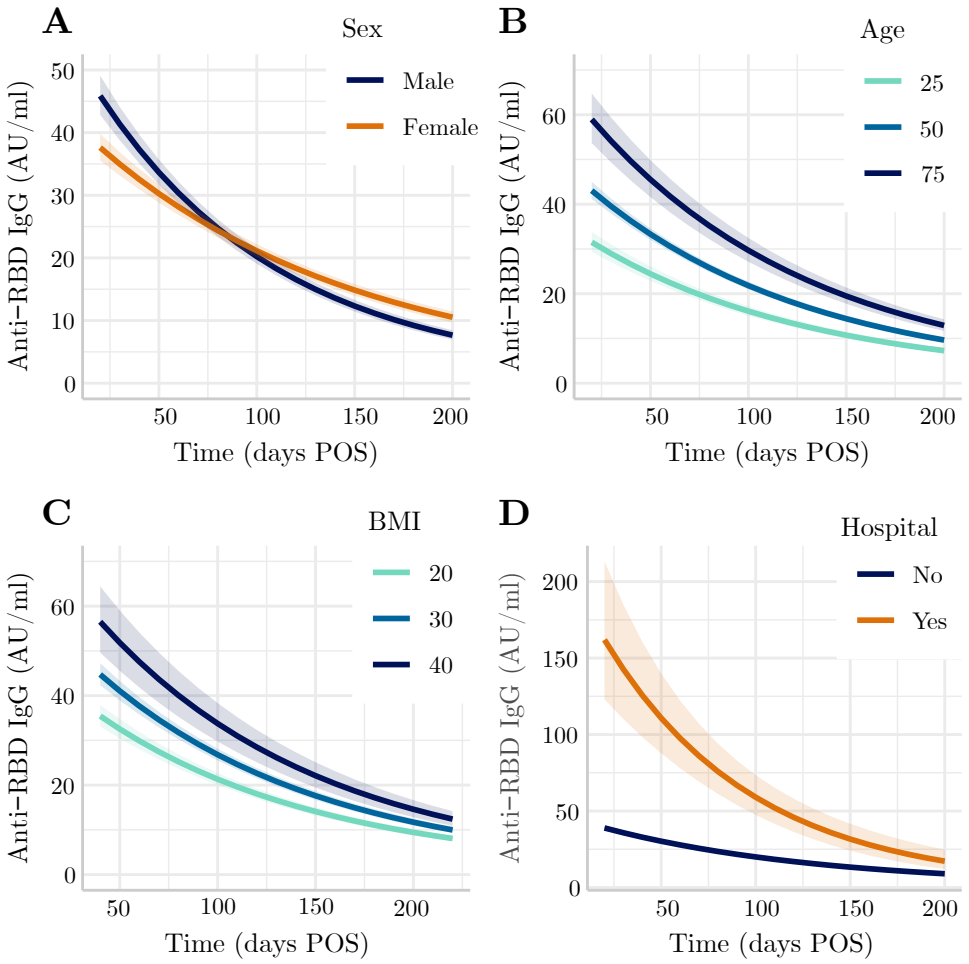

Figure 6.2: Associations between donor/clinical characteristics and antibody levels. The effects of variables (A) sex, (B) age, (C) BMI, and (D) hospital admission on predicted antibody decline. Note that age and BMI are included in the model as continuous predictors; for clarity, the associations are only plotted for three values. Light-coloured bands represent 95% confidence intervals.

Term	Estimate	95% CI
Random effects		
Intercept [log(peak IgG)]	2.382	2.274 – 2.490
Slope [delta log(IgG) per day]	-0.010	-0.011 – -0.101
Fixed effects on the intercept		
*Sex: female	-0.017	-0.063 – 0.096
Age (per 10 years)	0.128	0.100 – 0.157
BMI (per 5 points)	0.119	0.097 – 0.164
Hospital admission: yes	1.156	0.934 – 1.379
Headache: yes	-0.113	-0.193 – -0.032
Anosmia: yes	-0.111	-0.189 – -0.033
Nasal cold: yes	-0.101	-0.177 – -0.025
Dry cough: yes	0.095	0.019 – 0.171
Fatigue: yes	0.140	0.044 – 0.236
Diarrhoea: yes	0.148	0.043 – 0.252
Muscle weakness: yes	0.172	0.083 – 0.261
Shortness of breath: yes	0.196	0.111 – 0.280
Fever: yes	0.228	0.149 – 0.308
Fixed effects on the slope		
Sex: female	0.003	0.002 – 0.004
Hospital admission: yes	-0.004	-0.007 – -0.001

Table 6.2: Point estimates and 95% confidence intervals of fixed effects on log-transformed IgG levels.
* The effect of sex on the intercept (peak IgG) was not statistically significant, but the variable is not excluded due to its effect on the slope.

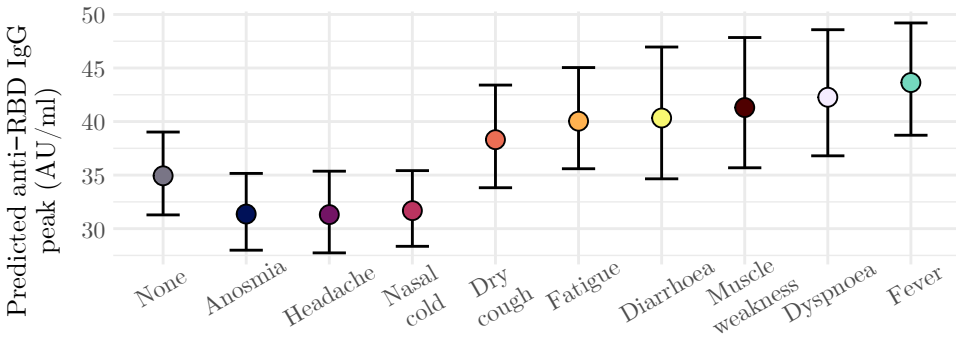

Figure 6.3: Predicted impact of various symptoms on anti-RBD IgG peak concentration. Estimated peak IgG concentrations when different symptoms are displayed. For each of the symptoms here, the difference in peak IgG as compared to the group without this symptom is statistically significant with $p < 0.001$.

The largest difference was found for the variable *hospital admission*. Donors admitted to the hospital had considerably higher antibody levels, with an estimated difference of 77.8 AU/ml on the peak IgG concentration. These donors also have a faster rate of antibody decay, corresponding to an estimated half-life of 48 days (95% CI: 40-58 days) for men, and 60 days (95% CI: 49-80 days) for women.

Variance explained by model

In the null-model that was fitted in step 1 (without any fixed effects), all variation in peak IgG and half-life was attributed to the individual variation per donor. As fixed effects were added in step 2, part of this variation was now explained by these fixed effects, and the variation explained by the random effects decreased. Table 6.3 shows the variance of the random effects per donor in the null-model, as well the variance of the random effects as after adding donor characteristics as covariates (step 2), and after adding the clinical information (step 3). The variance reduction relative to the null-model (step 1) by the addition of extra explanatory variables in each step is also provided. Model fit was compared using the Akaike Information Criterion (AIC) and tested for statistical significance using a nested ANOVA, results of which are shown in Table 6.3.

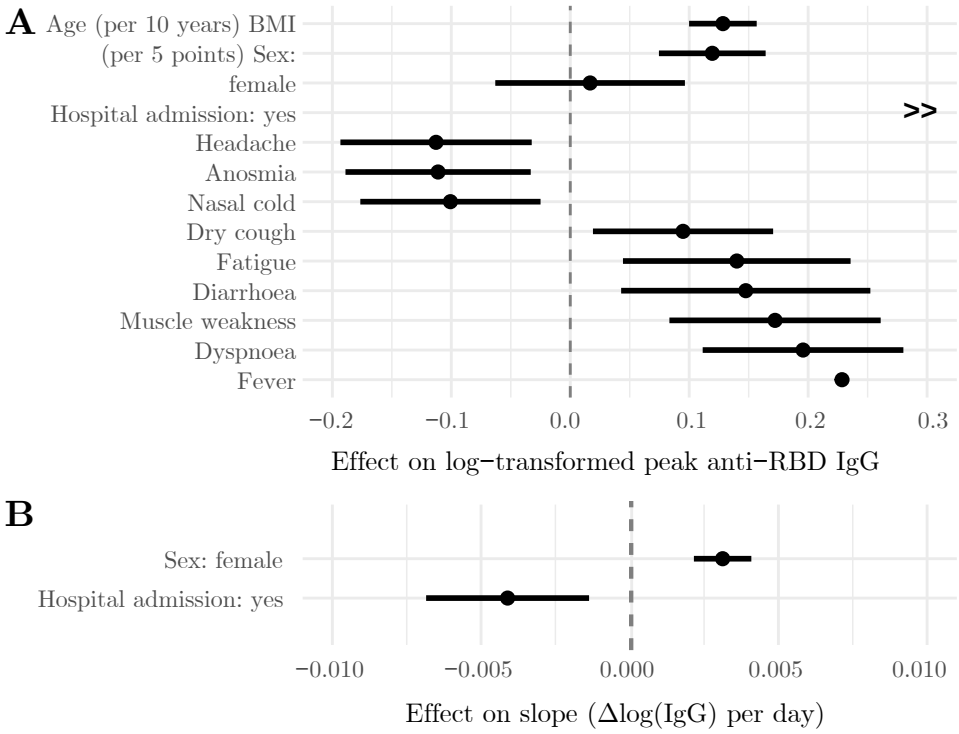

Figure 6.4: Effect size and 95% confidence intervals of fixed effects on anti-RBD IgG peak concentration (log-transformed) and the slope.

	Variance of random effect on peak log(IgG)	Variance of random effect on slope Δlog(IgG)/day	AIC
Step 1: null-model	0.8814	0.0497	11886
Step 2: donor characteristics	0.7758 (-12%)	0.0485 (-2.4%)	11615 ($p < 0.001$)
Step 3: donor characteristics + clinical information	0.6610 (-25%)	0.0481 (-3.2%)	11290 ($p < 0.001$)

Table 6.3: Variance of random effects in models of all three steps. Percentual variance decrease relative to the null-model is given in brackets. P-values are relative to the previous step, obtained with ANOVA.

Discussion

In this retrospective observational study, we investigated potential associations between SARS-CoV-2 specific antibody kinetics and various donor characteristics and COVID-19 symptoms. To our knowledge, this is currently the largest study that describes such associations. Individual antibody responses were modelled using a linear mixed-effects model, from which peak IgG concentration and antibody half-life were determined. Symptoms and donor characteristics were obtained from a questionnaire. Our study shows that the SARS-CoV-2 antibody response is associated with patient characteristics like sex, age, and BMI. Of note, we also found that specific COVID-19 symptoms are associated with antibody levels.

As reported earlier, we found a large variation in anti-RBD antibody peak levels. A strength of our study are the longitudinal measurements, which enabled us to reliably estimate the peak level of each individual donor independent on the timing of the first antibody measurement. Only a quarter of the variation in peak IgG concentration between patients can be explained by associations with donor characteristics and disease symptoms. To a lesser degree, donor characteristics were also associated with differences in antibody half-life, which was also variable between donors, albeit less than the peak level. The antibody half-life was not correlated to peak levels. Whether these differences in antibody half-life reflect differences in protection for re-infection will be investigated, and this thoroughly characterised donor cohort can serve as bench mark for those studies.

Six symptoms (dry cough, fatigue, diarrhoea, fever, dyspnoea, muscle weakness) were associated with higher IgG concentrations and three symptoms (headache, anosmia, nasal cold) were associated with lower peak IgG concentrations against RBD. This association between symptoms and antibody levels may possibly reflect the fact that the SARS-CoV-2 virus frequently initiates infection in the upper airways (mild symptoms and low IgG levels) before spreading through the body (severe symptoms and high IgG levels). Headache, anosmia and nasal cold were common symptoms, each present in at least 50% of patients in our population. Fatigue was present in more than 70% of patients and was associated with higher peak IgG concentration, suggesting more severe illness. A previous study in a hospital cohort found that fatigue and dyspnoea are prognostic for severe infection, and a stuffed nose (comparable to nasal cold) for mild infection, which is in line with our findings. [91]

Furthermore, we found higher age and BMI to be associated with higher peak IgG concentrations. Sex was not associated with peak IgG concentration, but men

had significantly shorter antibody half-lives than women (62 vs 72 days respectively). The small group of patients that had been admitted to hospital displayed both higher peak IgG concentrations and shorter half-lives. Probably this effect is the result of the presence of short-lived plasmablasts that produce high levels of antibodies. Previous studies found sex differences in COVID-19 immune responses, with higher IgG concentrations associated with male sex, older age, and hospitalisation. [92, 93, 94] Although our results are consistent with these findings for age and hospitalisation, we found that the association between male sex and higher peak IgG concentration was not significant after correction for age and BMI. This suggests that the previously found association with male gender was possibly due to the increased risk of severe disease in men. Most studies on differences in antibody response are performed in hospital cohorts, our study population consisted mainly of recovered patients that were not admitted to hospital (96.7%), and therefore disease severity is expected to be lower. Consistently, BMI in the non-hospitalised group was 25.9 compared to 28.8 in hospitalised patients.

A strength of our study is the large number of recovered patients included in our study population. The status of Sanquin as the only blood bank in the Netherlands, combined with well-established connections with municipal health services, allowed us to invite people with a positive PCR test to become CCP donors after recovery. This allowed us to both include non-hospitalised and hospitalised patients in the cohort. However, we could only include donors who were healthy enough to regularly donate plasma, which means that our results are mainly applicable towards patients with a mild outcome. As a result our study is more representative of the total COVID-19 patient population than studies on hospitalised patient cohorts. It should also be noted that some bias may be present in our data, as symptoms are self-reported by patients after recovery. Relatively mild symptoms, such as nasal cold, may therefore be underreported by patients who at the same time experienced more severe symptoms, such as fever or dyspnoea. However, this explanation is unlikely to negate the association we found, as all symptoms associated with lower peak IgG were present in more than 50% of patients.

In conclusion, our study indicates that several COVID-19 symptoms are associated with SARS-CoV-2 antibody levels in addition to the previously described association with sex, age, and BMI. Discovery of these associations aids us in understanding why antibody responses differ between patients. The predictive value of IgG concentrations could also be used by blood banks to pre-select individuals with high and/or stable antibody levels as potential CCP donors.

Appendix

Questionnaire anti-SARS-CoV-2 donors

Note: the original questionnaire was in Dutch, this is a translated version.

Q1 What is your donor ID? You can find this in the accompanying email.

Q36 What is your date of birth?

Day _____ - Month _____ - Year _____

Q2 What is your sex?

- ❍ Male
- ❍ Female

Q3 How would you describe your COVID-19 status?

- ❍ I suspect I have had COVID-19 because I have had a positive PCR test.
- ❍ I suspect I have had COVID-19 because antibodies have been detected in my blood.
- ❍ Other: _____

Q4 Where did you contract the infection?

- ❍ In the Netherlands
- ❍ Abroad

Q5 (if Q4 answered with 'abroad') In which country did you contract the infection?

Q6 Why were you tested for presence of the coronavirus?

- ❍ Because I was ill/had symptoms
- ❍ Because I was in contact with a (possibly) infected person
- ❍ Because of my occupation (health care, contact profession)
- ❍ Other: _____

Q7 Did you experience the following symptoms?

- ❑ Nasal cold/coryza
- ❑ Sore throat
- ❑ Dry cough
- ❑ Fatigue
- ❑ Sputum production
- ❑ Muscle or joint ache
- ❑ Headache
- ❑ Fever
- ❑ Shortness of breath
- ❑ Diarrhoea
- ❑ Nausea
- ❑ Vomiting
- ❑ Chills
- ❑ Sneezing
- ❑ Skin rash
- ❑ Feeling confused
- ❑ Muscle weakness
- ❑ Loss of/less smell or taste

Q8-25 (for each symptom answered with 'yes' in Q7) How much were you affected by this symptom?

- ❍ Very mildly affected
- ❍ Mildly affected
- ❍ Moderately affected
- ❍ Severely affected

Q37 Did you have pneumonia?

- ❍ Yes
- ❍ No

Q26 When did your symptoms start? If you don't remember exactly, please make an estimate.

Day _____ - Month _____ - Year _____

Q27 When did your symptoms end? If you don't remember exactly, please make an estimate.

Day _____ - Month _____ - Year _____

Q28 Were you admitted to hospital for these symptoms?

- ❍ Yes
- ❍ No

Q29 (if Q28 is 'yes') On which date were you admitted to hospital?

Day _____ - Month _____ - Year _____

Q30 (if Q28 is 'yes') Were you admitted to intensive care?

- ❍ Yes
- ❍ No

Q31 (if Q30 is 'yes') How many days were you in intensive care in total?

Q32 (if Q28 is 'yes') Were you given extra oxygen?

- ❍ Yes
- ❍ No

Q33 (if Q28 is 'yes') Did you receive artificial ventilation?

- ❍ Yes
- ❍ No

Q34 (if Q28 is 'yes') On which date were you discharged from the hospital?

Day _____ - Month _____ - Year _____

Q35 Are you in one or more of the following risk groups?

- ❑ People aged 70 or older
- ❑ People with chronic airway or lung disease and under treatment by pulmonologist
- ❑ Chronic heart disease patients under treatment by cardiologist
- ❑ People with diabetes that is not well regulated and/or with complications
- ❑ People with kidney disease who need dialysis or are waiting for a kidney transplantation
- ❑ People with lowered immunity to infection due to medication use for autoimmune disease
- ❑ People who have had an organ or stem cell transplantation
- ❑ People without a spleen or without a functioning spleen
- ❑ People with a blood disease
- ❑ People with lowered immunity due to immunity-lowering medication
- ❑ Cancer patients who have had chemotherapy and/or radiation in the past 3 months
- ❑ People with severe immune disorder that requires medical treatment
- ❑ People with HIV infection who are not (yet) under treatment, or with HIV infection with CD4 under 200/mm2
- ❑ People with severe liver disease
- ❑ People with a BMI over 40

Q38 If there is anything you would like to add, or explain an answer further, please do so here.

Supplemental figures and tables

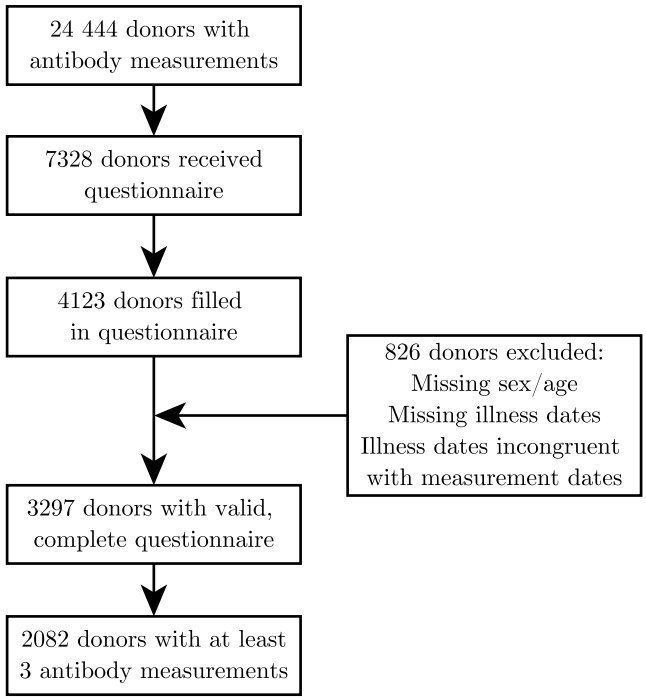

Figure S6.1: Flowchart showing the criteria for inclusion and exclusion of donors.

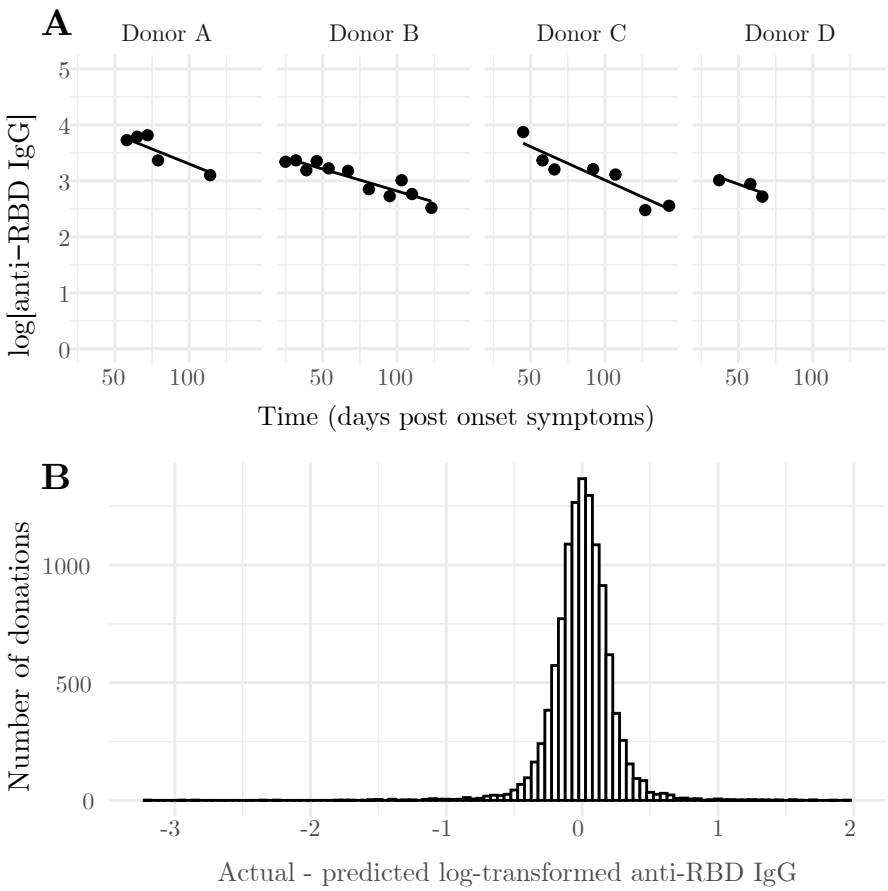

Figure S6.2: Null model fit (step 1). (A) Measured anti-RBD IgG levels (points) and fitted line as estimated by the linear model for four randomly selected donors, and (B) distribution of residuals over all observations, for all donors.

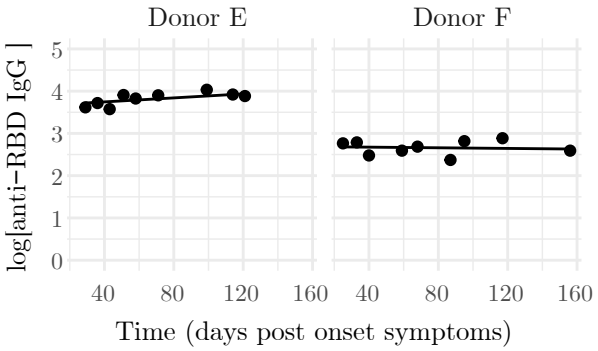

Figure S6.3: No decay in antibody levels. Example of a donor with increasing IgG levels (left panel) and one with near-constant IgG levels (right panel). Estimated slopes for these donors are 0.0024 and -0.0151, corresponding to estimated half-lives of -292 and 1843 days, respectively.

Fixed effect on intercept	Sum of squares	P-value
Weight	0.0070	0.762
Blood group ABO	0.4456	0.121
Height	0.3630	0.030
Blood group RhD	0.4638	0.014
BMI	3.888	<0.001 *
Age	8.752	<0.001 *
Fixed effect on slope	**Sum of squares**	**P-value**
BMI	0.002	0.890
Age	0.004	0.831
Height	0.009	0.735
Weight	0.035	0.500
Blood group ABO	0.188	0.483
Blood group RhD	0.085	0.294
Sex	3.403	<0.001 *

Table S6.1: Sum of squares and p-values of fixed effects after step 2, calculated by backward stepwise reduction

Fixed effect on intercept	Sum of squares	P-value
Sneezing	0.000	0.971
Vomiting	0.004	0.812
Confusion	0.011	0.700
Coughing up mucus	0.084	0.296
Throat ache	0.120	0.210
Joint/muscle ache	0.251	0.070
Nausea	0.247	0.072
Intensive care admission	0.274	0.059
Shivers	0.345	0.034
Skin rash	0.451	0.015
Anosmia	0.645	0.004 *
Fatigue	0.567	0.003 *
Nasal cold	0.449	0.002 *
Dry cough	0.394	0.002 *
Diarrhoea	0.462	0.001 *
BMI	2.000	<0.001 *
Hospital admission	8.555	<0.001 *
Fever	1.836	<0.001 *
Shortness of breath	1.510	<0.001 *
Fixed effect on slope	Sum of squares	P-value
Sex	3.470	<0.001 *

Table S6.2: Sum of squares and p-values of fixed effects after step 3, calculated by backward stepwise reduction.

CHAPTER

7

Explainable hemoglobin deferral predictions using machine learning models: interpretation and consequences for the blood supply

Published in: Vox Sanguinis 117(11): 1262-1270. doi:10.1111/vox.13350

Authors: M Vinkenoog, M van Leeuwen, MP Janssen

Abstract

Background - Accurate predictions of hemoglobin deferral for whole-blood donors could aid blood banks in reducing deferral rates and increasing efficiency and donor motivation. Complex models are needed to make accurate predictions, but predictions must also be explainable. Before the implementation of a prediction model, its impact on the blood supply should be estimated to avoid shortages.

Methods - Donation visits between October 2017 and December 2021 were selected from Sanquin's database system. The following variables were available for each visit: donor sex, age, donation start time, month, number of donations in the last 24 months, most recent ferritin level, days since last ferritin measurement, hemoglobin at nth previous visit (n between 1 and 5), days since the nth previous visit. Outcome hemoglobin deferral has two classes: deferred and not deferred. Support vector machines were used as prediction models, and SHapley Additive exPlanations values were used to quantify the contribution of each variable to the model predictions. Performance was assessed using precision and recall. The potential impact on blood supply was estimated by predicting deferral at earlier or later donation dates.

Results - We present a model that predicts hemoglobin deferral in an explainable way. If used in practice, 64% of non-deferred donors would be invited on or before their original donation date, while 80% of deferred donors would be invited later.

Conclusions - By using this model to invite donors, the number of blood bank visits would increase by 15%, while deferral rates would decrease by 60% (currently 3% for women and 1% for men).

Introduction

Sanquin, the Dutch national blood service, collects over 400 000 whole-blood donations from non-remunerated, voluntary blood donors every year. Women may donate a maximum of three times per year, and men five times. Hemoglobin levels are tested before every donation to prevent blood collection from donors with insufficient iron. The minimum hemoglobin level for blood donation is 7.8 mmol/L for women and 8.4 mmol/L for men; if the capillary hemoglobin test (HemoCue) shows a lower value, the donor is deferred for 3 months, that is, sent home without donating blood. If the hemoglobin value is more than 0.5 mmol/L below the donation threshold, the donor is referred to a donor physician. Additionally, since October 2017, ferritin levels have been measured in each new donor, as well as after every fifth donation in repeat donors. Donors are deferred for 6 months if their ferritin level is between 15 and 30 µg/L, or for 12 months if their ferritin level is below 15 µg/L. This ferritin deferral policy was implemented because hemoglobin is a poor indicator of iron stores, as iron deficient donors can still present with sufficient hemoglobin levels until the iron deficiency is very severe.

While it is important to defer donors that do not meet donation requirements, sending donors home without giving them the opportunity to donate is discouraging and costly. Previous studies have shown that donors are less likely to return to the blood bank after a deferral for low hemoglobin than after a successful donation, especially if it concerns their first blood bank visit. [28] This is less likely after deferral for low ferritin levels, which occurs by letter after the donation, indicating that post-donation deferral is less demotivating for donors than on-site deferral. [95] The implementation of ferritin testing has had a considerable impact on the blood supply, as a large part of the existing donor population (53% of women and 42% of men) were found to have ferritin levels below 30 µg/L and had to be deferred. [96] However, this has had the intended positive impact on donor deferral rates due to low hemoglobin, which decreased from 8% for women and 3% for men in 2016 to 3% for women and 1% for men in 2021. [97]

Although percentage-wise, hemoglobin deferral rates are quite low in the Netherlands, they still amount to about 8000 deferrals each year, and there is a risk of permanently losing these donors. To reduce deferral rates and improve donor motivation, we should re-think hemoglobin deferral policies. One tool that can be used for this purpose is a hemoglobin deferral prediction model. Many of these prediction models have already been developed, including models that predict personalised do-

nation intervals. [98, 99, 100] Prediction models can be used in the donor invitation process by predicting hemoglobin deferral for eligible donors and only inviting those donors that are predicted to not be deferred. Because deferred donors are only a small proportion of the total donor population, it has proven difficult to accurately identify them, and hence prediction models are not used in practice yet.

We present a novel machine learning hemoglobin deferral prediction model based on donor characteristics and donation history. New in our approach is that we use SHapley Additive exPlanations [101] to explain how the model uses the variables in its predictions and relate these explanations to known physiological processes. This gives valuable insight into the associations that are learned by the model; if prediction models are to be used to make decisions in practice, the user must understand how the model makes these decisions. Moreover, we show the potential impact that prediction models can have on the total blood supply, if these are to be used to guide donor invitations, by calculating deferral probabilities at multiple time points for each donor. By both explaining the predictions and assessing the impact of the model on the blood supply, we remove two important limitations that currently prevent blood services from implementing prediction models.

Methods

Data

Data on blood bank visits by whole-blood donors were extracted from Sanquin's database system eProgesa, for donations. Only data from donors who explicitly provided informed consent for the use of their data for scientific research were used. This consent is given by more than 99% of all donors. For each visit, the following information was collected: donor sex, donor age, donation date, donation (registration) time, hemoglobin level and ferritin level. Ferritin is measured at every new donor intake and upon every fifth donation in repeat donors. Therefore, ferritin levels are unavailable for most donations. By using these data, predictor variables were calculated for each visit, as described in Table 7.1.

In total, 938 710 blood bank visits (excluding new donor intakes and donation types other than whole blood) by 241 131 unique donors were registered between October 2017 and December 2021. After excluding visits for which no previous ferritin measurement was available, 458 615 blood bank visits by 157 423 unique donors remained for the analysis.

Variable	Unit or values	Description
Sex	male, female	Biological sex of the donor; separate models are trained for men and women
Age	years	Donor age at time of donation
Time	hours	Registration time when the donor arrived at the blood bank
Month	1–12	Month of the year that the visit took place
NumDon	count	Number of successful (collected volume >250 ml) whole-blood donations in the last 24 months
FerritinPrev	µg/L	Most recent ferritin level measured in this donor
DaysSinceFer	days	Time since this donor's last ferritin measurement
HbPrevn	mmol/L	Hemoglobin level at nth previous visit, for n between 1 and 5
DaysSinceHbn	days	Time since related hemoglobin measurement at nth previous visit, for n between 1 and 5

Table 7.1: All predictor variables used in the prediction models.

The outcome variable *HbOK* is dichotomous; deferral (hemoglobin level below the eligibility threshold for donation) or non-deferral (hemoglobin equal to or above the threshold).

Analyses

Support vector machines (SVMs) [102] are used to predict hemoglobin deferral. SVMs are supervised machine learning models that find the optimal hyperplane separating the outcome classes based on the predictor variables of a so-called training set. After fitting the model on the training set, the model can predict the outcome class of unseen observations called the test set. It also gives the probability of an observation belonging to each outcome class. We chose SVMs as a classification algorithm because all predictor variables are numeric, and it is computationally less expensive than, for instance, K-nearest neighbours or (dynamic) linear mixed models.

For each sex, five SVMs were trained, named SVM-n for n between one and five, indicating the number of previous blood bank visits (*HbPrevn* and *DaysSinceHbn*) used as predictor variables. Donors are only included in SVM-n if they have at least n previous visits; therefore, sample sizes decrease from SVM-1 to SVM-5. Blood bank visits before 2021 were used as the training set, while visits in 2021 were used as the test set to validate performance on unseen data. This division was chosen over a random training/test division because if these models were used in practice, they would be

Metric	Outcome class	Definition
Precision	Deferral	The proportion of donations correctly classified as deferrals by the model, out of all donations classified as deferrals.
Recall	Deferral	The proportion of donations correctly classified as deferrals by the model, out of all donations classified as true deferrals.
Precision	Non-deferral	The proportion of true non-deferrals, out of all predicted non-deferrals.
Recall	Non-deferral	The proportion of predicted non-deferrals, out of all true non-deferrals.

Table 7.2: Interpretation of performance metrics.

trained on all historical data and applied to future data. We used a paired t-test to assess the difference in deferral rates between training and test sets of donors of the same sex with the same number of previous donations. To assess the generalisability of the model to new donors, we did a separate experiment in which the test set is comprised of the last blood bank visit of 20% of all unique donors, and the training set includes all donations from the remaining 80% of donors.

For each of the 10 models, that is, SVM-1 through SVM-5 for both sexes, hyperparameters were optimised separately, using stratified (on the outcome variable) five-fold cross-validation within the training set data (and thus not using the test data). Hyperparameters were optimised using grid search, using balanced accuracy as a scoring method, defined as the weighted average of recall in both classes (see Table 7.2 for the definition of recall). This method is especially suitable for imbalanced datasets because it uses class-balanced sample weights to determine the average recall.

Precision and recall were determined and compared for training and test datasets for each model. Both metrics are calculated for both outcome classes. A practical interpretation of these metrics is given in Table 7.2.

To explain the model predictions, we used SHapley Additive exPlanations (SHAP) values, a model agnostic explainer. SHAP values show the contribution of each variable to the prediction for each individual observation, which is even more informative than coefficients returned by, for example, linear models. By summarizing observation-based contributions, we obtain variable importance measures for a model that does not have interpretable coefficients.

Potential impact on the blood supply

We assessed the potential impact of using SVMs to guide donor invitations by predicting deferral for all blood bank visits that took place in 2021 (the test set). For each observation, we used information of all previous blood bank visits (up to five) available as predictor variables. This means that SVM-1 is used when only one previous visit is available, SVM-2 if there are two previous visits, etc.

If prediction models are to be used in practice, they should estimate the deferral probability for different days in the future and invite a donor for the first occurrence where the non-deferral probability would exceed a preset value. To simulate this, we predicted hemoglobin deferral each week from 1 year before the original donation date to 1 year after, by adjusting all time-related variables. If the predicted donation interval were to be less than the minimum donation interval (57 days for men, 122 days for women), the latter would be applied.

We compare all original donation intervals with the donation intervals as proposed by the model. Dividing the sum of the original donation intervals by the sum of the model-guided donation intervals gives the relative change in blood bank visits per time unit and hence the relative yield of blood donations.

Software

All analyses were performed in Python 3.9, using modules numpy [103] and pandas [104] for data processing, sklearn [105] for model training and predictions, shap [101] for calculating SHAP values, and matplotlib [106] for creating graphs. The analysis code is available as a GitHub repository and indexed on Zenodo at https://doi-org.ezproxy.leidenuniv.nl/10.5281/zenodo.6938112.

Results

Table 7.3 shows the sample sizes of training and test datasets for each model. Deferral rates in the training datasets are 3.19% (SD 0.28) for women and 1.22% (SD 0.09) for men; in the test sets, they are 3.42% (SD 0.24) for women and 1.21% (SD 0.08) for men. Using a paired t-test, the difference in deferral rate between the training and test datasets is significant for women ($p = 0.002$) but not for men ($p = 0.070$). No correction was made for the differing deferral rates, as the models are intended for future predictions, and in practice, the deferral rate of future blood bank visits is unknown. Also, a change in deferral rate should be correctly predicted by the model if

Model	Training		Test	
	Women	Men	Women	Men
SVM-1	128 173 (4084; 3.19%)	121 746 (1339; 1.10%)	110 372 (3696; 3.35%)	98 324 (1074; 1.09%)
SVM-2	83 532 (2884; 3.45%)	96 441 (1133; 1.17%))	85 131 (3065; 3.60%)	84 000 (984; 1.17%)
SVM-3	59 720 (2032; 3.40%)	79 690 (997; 1.25%)	67 167 (2451; 3.65%)	72 576 (902; 1.24%)
SVM-4	47 317 (1494; 3.16%)	67 934 (887; 1.31%)	54 090 (1874; 3.46%)	63 447 (806; 1.27%)
SVM-5	40 604 (1113; 2.74%)	59 611 (768; 1.29%)	45 208 (1378; 3.05%)	55 582 (699; 1.26%)

Table 7.3: Sizes of training and test datasets per model. The number and percentage of deferrals is given in brackets.

the mechanism causing this change can be learned from the data. Deferral rates differ between models due to small differences in the data between subsets of the data (see Table 7.4). This is not a problem as long as the same associations between predictor variables and outcome are found in all subsets of the data, which is described in the feature importance part of the results.

Although the training datasets consist of 3 years of data, and the test datasets of only 1 year, their sizes are similar and sometimes the test dataset is even larger. This is because donations are only included from donors for whom at least one ferritin measurement was available. As ferritin screening was implemented using a stepped wedge approach (the first blood bank locations started in October 2017, but only in November 2019 all locations were included), the number of donors that could be included in the training dataset was limited. [97]

Marginal distributions of predictor variables are described in Table 7.4. As the number of previous donations increases, the median age increases from 30 to 36 years for women and from 34 to 38 for men. The median values of the last ferritin measurement decreased from 47 µg/L in SVM-1 to 39 µg/L in SVM-5 for women and from 77 to 47 µg/L for men. The median time between consecutive donations increases from SVM-1 to SVM-5, while previous hemoglobin levels are consistent across models, as well as different numbers of previous visits.

			Women		
Previous visits	≥ 1	≥ 2	≥ 3	≥ 4	≥ 5
Age	30 (23–47)	32 (24–48)	34 (25–50)	35 (26–51)	36 (37–52)
NumDon	1 (0–3)	2 (1–3)	3 (2–4)	3 (2–4)	3 (3–4)
FerritinPrev	47 (33–74)	46 (33–70)	44 (32–65)	41 (31–59)	39 (29–55)
DaysSinceFer	237 (125–420)	329 (197–497)	383 (260–547)	400 (230–572)	372 (204–567)
HbPrev1	8.5 (8.1–8.9)	8.5 (8.1–8.9)	8.5 (8.1–8.9)	8.5 (8.1–8.9)	8.5 (8.1–8.9)
DaysSincePrev1	135 (105–196)	154 (132–211)	158 (132–217)	167 (133–224)	173 (133–236)
HbPrev2		8.5 (8.1–8.9)	8.5 (8.1–8.9)	8.5 (8.1–8.8)	8.5 (8.1–8.8)
DaysSincePrev2		302 (255–412)	328 (271–445)	336 (273–468)	349 (280–493)
HbPrev3			8.5 (8.1–8.8)	8.4 (8.1–8.8)	8.4 (8.1–8.8)
DaysSincePrev3			482 (398–644)	511 (420–674)	528 (430–696)
HbPrev4				8.4 (8.1–8.8)	8.4 (8.1–8.8)
DaysSincePrev4				674 (553–871)	709 (581–904)
HbPrev5					8.4 (8.1–8.8)
DaysSincePrev5					877 (721–1107)
			Men		
Previous visits	≥ 1	≥ 2	≥ 3	≥ 4	≥ 5
Age	34 (26–48)	35 (27–49)	36 (27–50)	37 (28–51)	38 (28–51)
NumDon	3 (1–5)	4 (2–5)	4 (3–6)	5 (3–6)	5 (4–6)
FerritinPrev	77 (44–141)	66 (40–126)	57 (38–108)	52 (36–89)	47 (35–73)
DaysSinceFer	200 (100–335)	232 (151–365)	257 (177–378)	271 (186–385)	267 (173–387)
HbPrev1	9.4 (9.0–9.9)	9.4 (9.0–9.9)	9.4 (9.0–9.8)	9.4 (8.9–9.8)	9.4 (8.9–9.8)
DaysSincePrev1	81 (63–133)	90 (67–147)	92 (69–160)	98 (70–168)	105 (70–176)
HbPrev2		9.4 (9.0–9.8)	9.4 (9.0–9.8)	9.4 (8.9–9.8)	9.4 (8.9–9.8)
DaysSincePrev2		185 (128–287)	196 (147–302)	210 (153–315)	219 (158–330)
HbPrev3			9.4 (9.0–9.8)	9.4 (9.0–9.8)	9.4 (8.9–9.8)
DaysSincePrev3			302 (225–441)	322 (238–463)	335 (245–485)
HbPrev4				9.4 (8.9–9.8)	9.4 (8.9–9.8)
DaysSincePrev4				424 (315–600)	444 (330–620)
HbPrev5					9.4 (8.9–9.8)
DaysSincePrev5					552 (416–752)

Table 7.4: Marginal distributions of predictor variables, represented by median and interquartile ranges.

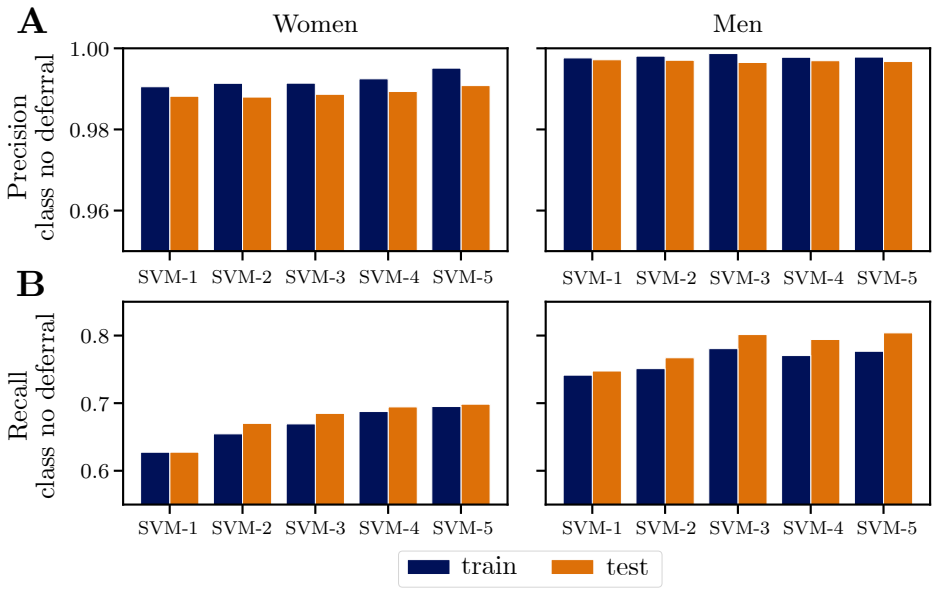

Figure 7.1: Performance metrics for all models. (A): Precision of class non-deferral; the proportion of successful donations among all predicted non-deferrals. The complement of the precision is the deferral rate, should the model be used to guide invitations. (B): Recall of class non-deferral; the proportion of successful donations that are predicted correctly. The complement of the recall is the proportion of missed donations, should the model be used to guide invitations. Note that the y-axes in are zoomed in to highlight the differences between various models.

Accuracy and model fit

Figure 7.1 compares precision and recall for class non-deferral across all models. Performance on the training and test sets are similar, indicating that the models are well-fitted. Both precision and recall increase as more previous blood bank visits are used to make predictions. Re-running all models only on donors with at least five previous blood bank visits did not change this observed increase in performance. The models handle the difference between the proportions of deferral in the training and test set very well: comparing the observed difference in deferral proportion in the training and test set to the predicted difference, the mean difference of these differences is only 0.05 percentage points (maximum: 0.12 percentage points). This indicates that the models are robust against (modest) changes in deferral rates.

Sex	Metric	Time split	Random split	Difference
Women	Precision	0.991	0.994	-0.003
	Recall	0.698	0.701	-0.003
Men	Precision	0.997	0.996	+0.001
	Recall	0.804	0.791	+0.013

Table 7.5: Precision and recall for outcome class non-deferral, compared between two different training/test splits.

Performance on a test set of unseen donors

Precision and recall for both outcome classes are similar for the different types of splits in training and test set. Table 7.5 shows the comparison in performance between the time split and the random split, as described in the methods section. Metrics are shown for SVM-5; the differences are smaller for all other models. For women, the random split has a higher precision and recall than the time split. For men, this is the other way around. For both sexes, the differences are minimal.

Feature importance and explanation of predictions

SHAP values were computed based on a random subset of 100 donations in the test set. Figure 7.2 shows the SHAP summary plot for the SVM-5 models, the summary plots for the other eight models are included in the online supplement of the published paper.

For all models, the most important predictor variable is the previous hemoglobin measurement (*HbPrev1*), and in general, more recent measurements are more important than earlier ones. The time since the previous hemoglobin measurements also ranks high on feature importance, but their chronological order is less well-preserved than the *HbPrev* variables.

The association between the feature value and impact on the prediction is as expected for most variables. For hemoglobin measurements, higher values are associated with predicted non-deferral. For *DaysSinceHb*, longer times since the previous hemoglobin measurement are indicative of predicted non-deferral. However, *DaysSinceHb4* shows the opposite association, meaning that when the fourth previous measurement was long ago, the chance of predicted non-deferral becomes lower, while higher would be expected.

Variable *NumDon* has the expected impact on prediction in all models but SVM-5 for female donors; in all other models, a higher number of recent donations shifts

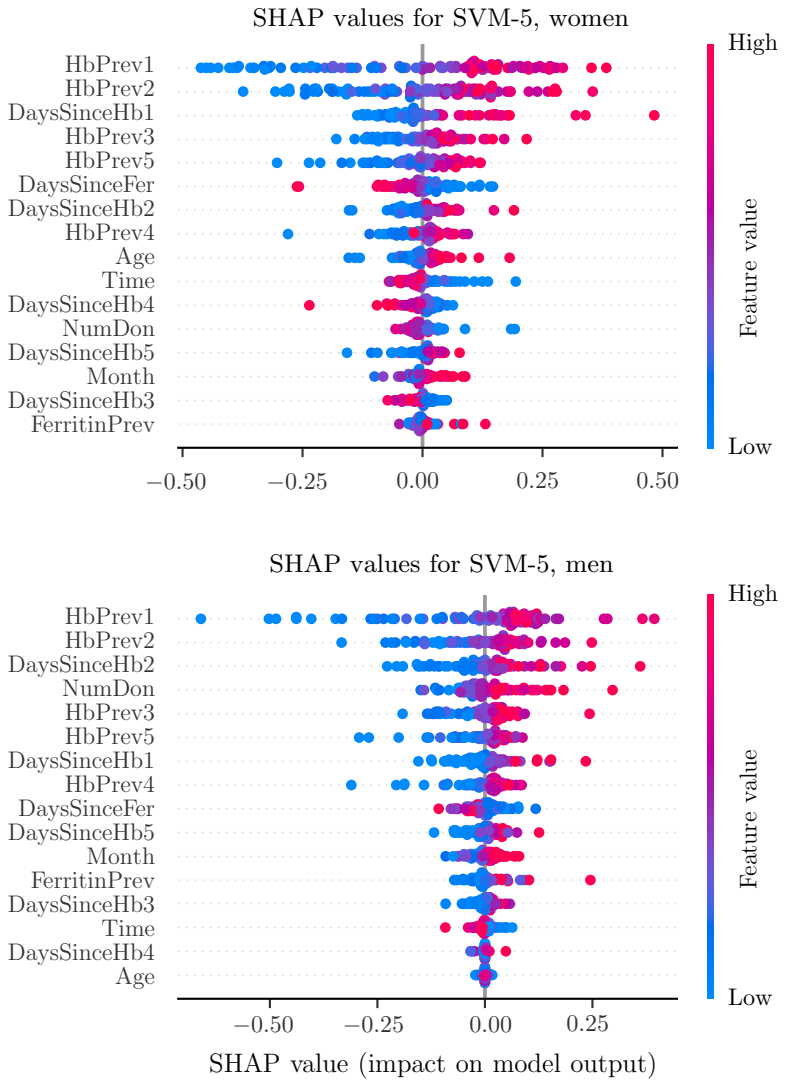

Figure 7.2: SHAP summary plots for predictions made by SVM-5, on 100 random donations from the test set. Each point represents one single observed donation. The location on the x-axis indicates the contribution of the predictor variable on the prediction (positive value: indicative class non-deferral, negative: indicative of class deferral) while the colour of the point indicates the relative value of the feature in that observation. The features on the y-axis are ordered by their relative importance, measured as the mean absolute SHAP value.

the prediction towards deferral. In most models, the number of donations is a more important predictor for men than for women, and it is always less important than all *HbPrev* variables.

The variable *FerritinPrev* shows the same association with the prediction as *HbPrev* variables: higher ferritin levels are associated with predicted non-deferral. Ferritin is a more important predictor for men than for women. For both sexes, the time since the previous ferritin measurement is more important than the actual ferritin level, and a higher value for *DaysSinceFer* makes predicted deferral more likely.

We know that for women, higher age makes deferral less likely (due to menopause), and the SHAP values confirm this relation. For men, age is one of the least important predictors, and there is no clear direction of the relation. The month of donation is of medium importance for both sexes, with predicted deferral being more likely earlier in the year. This captures the seasonal effect of temperature on hemoglobin as measured by the HemoCue. Donating earlier in the day (i.e., a lower value for variable *Time*) increases the likelihood of predicted non-deferral, which is supported by previous research showing that hemoglobin levels are highest in the morning and decrease throughout the day. [107]

Impact on blood supply

Figure 7.3 shows the cumulative count of donors as invited by the models relative to their original donation date. Once the model predicts non-deferral, it never predicts deferral at a later date. Of non-deferred donors, 50% would be invited more than 2 weeks earlier by the model, and 26% within 2 weeks from around the original donation date. Only 5% would not be invited within a year, causing a successful donation to be missed. Of deferred donors, only 13% would be invited earlier, while 40% would be invited over 3 months later. 28% would not be invited within 1 year. The majority of donors would be invited around their original donation date. For many donors, the original donation date was shortly after the minimum donation interval had passed, and as such, there was no room to invite them earlier.

Because the true hemoglobin level of donors on days other than their original donation date is unknown, we must make assumptions about the accuracy of the predictions in order to calculate a hypothetical number of donations and deferrals. In the most optimistic scenario, all donors who were not deferred on their original donation date would also not be deferred if they were invited earlier; and all donors who were deferred on their original donation date but are invited later by the model would not be deferred by then. In that scenario, only 5% of successful donations

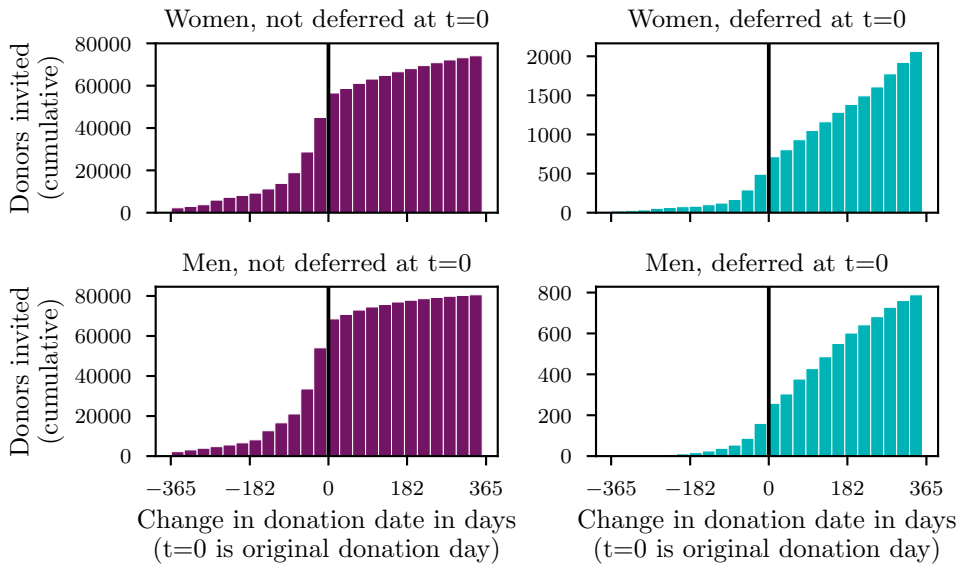

Figure 7.3: Cumulative distribution of the timing of donor invitations on basis of first predicted hemoglobin level above the donation threshold relative to the original donation date.

would be lost because the model would (incorrectly) not invite those donors, while the deferral rate would decrease by 60% (from 3% to 1% for women and from 1% to 0.4% for men).

We estimate the impact on the blood supply by comparing the length of the original donation interval to the donation interval as suggested by the model. For women, the median time between two donations decreases from 157 to 127 days using the prediction model. For men, the median time decreases from 92 to 63 days. Therefore, the total number of blood bank visits per time unit would increase by a maximum of 15%. This assumes that all donors who responded to the original invitation would also respond to the invitation if it would be sent at an earlier or later date. We also assume that all donors visit the blood bank within 1 week of the invitation. With the original invitations, 15% of donors that responded to the invitation visited the blood bank within 8 days, so the 15% increase in visits is likely to be a small overestimation. These assumptions may not hold for mobile donation sites but are reasonable for all regular donation sites, where 95.3% of all visits in our data occurred.

Discussion

This study presents an explainable machine learning approach to predict hemoglobin deferral in whole-blood donors using the information on previous donations and various donor characteristics. We show that we can prevent up to 60% of on-site low hemoglobin deferrals using the model to guide donor invitations.

To our knowledge, this is the first model using machine learning for explainable hemoglobin deferral prediction. An explainable model outcome is crucial for prediction models that are to be used in the context of a decision-support system concerning humans. SHAP values show that our models are able to learn biologically sensible associations. They support findings from other prediction models that found the previous hemoglobin value to be the best predictor for future deferral. We add to this by showing that including more previous donations will improve these predictions.

Although most associations found by SHAP values can be explained biologically, some seem to be caused by organisational policies. Higher values for *DaysSinceFer* are associated with predicted deferral; the opposite association is found for *DaysSinceHb* variables. For donors with fewer than five donations since the start of ferritin testing, the only ferritin measurement is the one taken at their new donor intake, and therefore the time since that previous ferritin measurement is equal to the time since their new donor intake. It is known that deferral becomes more likely once a donor has been

donating for a longer period of time.

The precision of class deferral is low, meaning that the predicted deferral is wrong for a substantial proportion of donors. However, by predicting deferral for different timepoints, we see a clear difference between deferred and non-deferred donors: non-deferred donors are in many cases invited earlier than their original donation date by the model, while deferred donors are mostly invited later or not at all, thereby reducing the deferral rate. In non-deferred donors, the median donation interval becomes shorter if invitations were guided by the model, and thus the number of blood bank visits per time unit would increase.

We can only calculate the accuracy of deferral predictions on the original donation date, as hemoglobin levels on other days are unknown. As hemoglobin levels slowly increase after a donation, non-deferred donors would also not be deferred if they were invited later. If they are invited earlier, we cannot know if their hemoglobin level is already above the deferral threshold. The same applies to deferred donors that are invited later by the model - it is plausible that their hemoglobin levels are above the threshold then, but not certain. Based on accuracy measures of predictions on the original donation dates, we can be fairly confident that the predictions are reliable.

Incorporating prediction models in hemoglobin deferral policies could bring many benefits to blood banks, but it is important to think about how they should be used. If the model is used in practice, the change in policy will lead to changes in the data. Models would therefore need updating by re-training on a regular basis. Additionally, it would be wise not to outsource invitations to the model completely, as that would hinder the model's ability to learn from its mistakes. Although deferrals incorrectly predicted to be non-deferrals would be discovered, we would never know how many donors were incorrectly not invited. This can be prevented by sending part of the invitations without using the model's predictions. In addition to using the model to predict deferral outcomes, the model can also be used to return a deferral probability, allowing blood banks to incorporate this probability in their risk assessment when inviting donors.

Our model is limited to predictor variables that are presently collected by Sanquin. Additional variables could be considered to improve prediction accuracy. Donor height and weight (optionally BMI or total blood volume), as well as smoking status, are examples known to be related to iron levels and are relatively easy to be included. Information on iron-related genetic markers or donor diet may also improve accuracy but are expensive to collect.

Based on the results of this study, we conclude that using prediction models to

guide donor invitations would bring multiple advantages to blood banks: lower deferral rates combined with shorter donation intervals would result in motivated and healthy donors, as well as a steady blood supply.

CHAPTER

8

An international comparison of hemoglobin deferral prediction models for blood banking

Published in: Vox Sanguinis 118(6): 430-439. doi:10.1111/vox.13426

Authors: M Vinkenoog, J Toivonen, T Brits, D de Clippel, V Compernolle, S Karki, M Welvaert, A Meulenbeld, K van den Hurk, J van Rosmalen, E Lesaffre, M Arvas, MP Janssen

Abstract

Background - Blood banks use a hemoglobin threshold before blood donation to minimise donors' risk of anemia. Hemoglobin prediction models may guide decisions on which donors to invite, and should ideally also be generally applicable, thus in different countries and settings. In this paper, we compare the outcome of various prediction models in different settings and highlight differences and similarities.

Methods - Donation data of repeat donors from the past 5 years of Australia, Belgium, Finland, the Netherlands and South Africa were used to fit five identical prediction models: logistic regression, random forest, support vector machine, linear mixed model and dynamic linear mixed model. Only donors with five or more donation attempts were included to ensure having informative data from all donors. Analyses were performed for men and women separately and outcomes compared.

Results - Within countries and overall, different models perform similarly well. However, there are substantial differences in model performance between countries, and there is a positive association between the deferral rate in a country and the ability to predict donor deferral. Nonetheless, the importance of predictor variables across countries is similar and is highest for the previous hemoglobin level.

Conclusions - The limited impact of model architecture and country indicates that all models show similar relationships between the predictor variables and donor deferral. Donor deferral is found to be better predictable in countries with high deferral rates. Therefore, such countries may benefit more from deferral prediction models than those with low deferral rates.

Introduction

To avoid blood donations by donors at risk of becoming anaemic, blood banks test the donors' hemoglobin (Hb) levels. In case of pre-donation testing, a low hemoglobin level leads to on-site deferral, which is demotivating for donors and makes them less likely to return to the blood bank than non-deferred donors. [29, 28] Additionally, it is in the interest of blood banks to keep deferral rates low to save time and costs. The ability to accurately predict low hemoglobin deferral and adjust donation intervals based on these predictions likely decreases deferral rates. In the last 15 years, various hemoglobin deferral prediction models, such as multiple logistic regression models, [99] Bayesian linear mixed models (LMM) [100, 108] and ensemble models, [98] have been evaluated by blood banks. Most prediction models use donors' previous hemoglobin measurements in combination with donor characteristics such as age and sex, but the prediction accuracy has been modest. Nonetheless, even models with modest accuracy could be beneficial in practice. [108] Accurate prediction of hemoglobin levels and/or deferral remains a difficult task, as many factors affect hemoglobin, and both intra- and inter-individual variation is large. Therefore, it stands to reason that machine learning models might improve the prediction accuracy over the traditional regression models, as they are capable of learning more complex associations between predictors and outcome variables. Support vector machines (SVMs) have been shown to predict hemoglobin deferral in Dutch donors reasonably well, [109] as do random forests (RFs) in Finnish donors. [108]

Most prediction models are developed and validated on donation data of a single country. [99, 98] Between countries, sets of available predictor variables differ widely. Ferritin levels, genotyping data, smoking status and iron supplementation are examples of variables that are associated with hemoglobin levels but are not systematically measured or recorded by most blood banks. [110] Therefore, prediction models using such variables cannot be applied to data from other blood banks. Additionally, differences in blood bank policies regarding donor deferral require models to be calibrated for each country separately.

The SanguinStats group is a collaboration of statisticians and epidemiologists from several countries carrying out research in the area of donor health. It currently consists of researchers from blood banks in Australia, Belgium, Denmark, Finland, the Netherlands, South Africa and the United Kingdom, as well as researchers with statistical expertise who are associated with research institutes other than blood banks. The aim of the SanguinStats group is to combine the available expertise and data sources to

develop and evaluate the outcome of state-of-the-art models in various settings.

In this first joint paper, we present a comparison of various hemoglobin deferral prediction models on data from five blood banks. The goal of this research is not to create the best performing predictor, but rather to use exactly the same models for all datasets and to compare the performance and importance of variables between countries. Therefore, only basic predictor variables that are available in all individual countries are included in the models. Comparing the importance of variables between countries will show whether models show the same relationships between the variables and hemoglobin deferral.

This is the first study to compare multiple hemoglobin deferral prediction models on datasets from multiple countries. The results can be used by other blood banks to anticipate benefits from collecting additional measurement data and the use of various predictors for the prediction of donor deferral.

Methods

Data sources and variables

Within each country, data were extracted from the blood banks' database, selecting data from whole blood donors from the past 5 years. The exact years differ per country because of the availability of up-to-date datasets. For each country, the timeframe of data collection was carefully selected to minimise iron-related blood bank policy changes in the dataset. In Australia, Finland and the Netherlands, there is one national blood bank (Australian Red Cross Lifeblood, Finnish Red Cross Blood Service and Sanquin Blood Bank, respectively), and data from these blood banks were used. In Belgium, data from Red Cross Flanders were used, which covers the whole of Flanders. In South Africa, data from South Africa National Blood Service were used, which is the major blood bank in the country.

For this study, only donors with five or more donation attempts were included to balance the trade-off between prediction accuracy (which has been shown to decrease with shorter time series at least in LMM) and data availability, as data becomes scarcer with higher thresholds of minimum donation numbers. [108]

The following donation-level variables are used in the prediction models:

- Donor age (*Age*)
- Days to previous donation (*Days to previous whole blood donation*)
- Time of day at the start of the donation (*Time*)
- Hemoglobin level at first donation (*First Hb*) (not used by dynamic linear mixed model [DLMM])
- Hemoglobin level at previous donation (*Previous Hb*) (not used by linear mixed model [LMM])
- Low hemoglobin at previous donation (*Previous visit low Hb*)
- Warm season (April–September for Northern hemisphere and October–March for Southern hemisphere) (*Warm season*)
- Number of consecutive deferrals since previous successful donation (*Consecutive deferrals*)
- Number of successful donations in last 5 years (*Recent donations*)
- Number of low hemoglobin measurements in the last 2 years (*Recent low Hb*)

Models were fitted separately for male and female donors. Unless otherwise specified, the analyses presented in this study were performed on a random subset of 10 000 donors per sex, to prevent differences in model performance between countries due to different dataset sizes. The outcome is a dichotomous variable: deferral or non-deferral.

Statistical methods

Five prediction models were compared in this study: a baseline model, random forest (RF), support vector machine (SVM), linear mixed model (LMM) and dynamic linear mixed model (DLMM). Note that these models are fundamentally very different. Each of the models is briefly described below.

The baseline model is a simple logistic regression model that estimates the likelihood of deferral as a function of only the hemoglobin level at the previous donation.

Random forest is a classification algorithm that consists of several decision trees, fitted on sub-samples of the data. It uses averaging to improve predictive accuracy

and prevent overfitting. The prediction output of an RF is the class selected by the majority of the decision trees. The RF takes as input all predictor variables listed in the previous section.

Support vector machine is a classification algorithm that aims to find the best hyperplane to separate both outcome classes in a multi-dimensional space. The SVM again takes all predictor variables listed in the previous section as input. Note that none of the three models mentioned above explicitly models the subsequent donations, but rather uses aggregated information on donation history (see list above). This is where these differ from LMM and DLMM, which include a donor-specific intercept as the only random effect.

The linear mixed model does not include previous hemoglobin as a predictor, but instead uses the first hemoglobin level. The dynamic linear mixed model, however, does include the previous hemoglobin as a predictor. Both LMM and DLMM are regression models that predict not hemoglobin deferral but the actual hemoglobin level. If this predicted hemoglobin level is lower than the country-specific donation threshold, deferral is predicted. These LMMs were trained in a Bayesian setting with weakly informative conjugate priors. They are described in more detail in a previous article [108], and they are essentially simplified versions of the models proposed by Nasserinejad et al. [100], excluding the modelling of the temporary reduction in hemoglobin after blood donation.

Model performance is assessed using the area under the precision–recall (AUPR) curve. As no perfect model exists, each model provides an estimate of the probability of deferring a donor. Depending on the probability that is applied as a classification threshold (so anyone with a higher probability of deferral is labelled *deferral* and the others *non-deferral*), a different number of correct and incorrect predictions will be found. The precision–recall curve is a graph in which the recall versus the precision of a prediction model at varying classification thresholds is shown, where precision is the proportion of correctly predicted deferrals of all predicted deferrals and recall is the proportion of all deferred donors that were correctly labelled as such. The higher the AUPR curve, the better the prediction model's performance. To fairly compare AUPR across countries, we adjusted the AUPR values by subtracting the countries' deferral rate. The adjusted value now indicates the improvement by the model over always predicting non-deferral.

SHapley Additive exPlanations (SHAP) values were used to quantify the contribution of each predictor variable to the prediction for each individual observation. [111] Because SHAP values are model-agnostic, they can be calculated and compared for

each model. This results in variable importance measures even for models that do not have interpretable coefficients, such as RF and SVM.

Docker container

To ensure that all collaborators perform exactly the same analyses, but without having to export data outside of their organisation or between jurisdictions, we implemented all models for hemoglobin deferral prediction in a Docker container whose development was started earlier. [108] The Docker platform is easy to install on all major operating systems. After installation, the Docker container image can be downloaded and the user can run all models presented in this paper in a secure environment (without requiring an internet connection). For this study, we added an implementation of the SVM to the container, in addition to some specific improvements to facilitate the comparison of outputs. Both the ready-to-use container image and its source code are freely available through Dockerhub and Github, respectively. All analyses presented in this paper were obtained using version 0.32 of the container. Analyses of the results were performed using the R language and environment for statistical computing [112], using packages dplyr [113] and tidyr [114] to handle data, and ggplot2 [115] to create graphs.

Variable	Australia	Belgium	Men Finland	Netherlands	South Africa
Number of donors	10 000	8552	10 000	10 000	10 000
Age in years	41 (29–54)	39 (25–52)	53 (41–60)	52 (39–60)	44 (33–54)
Mean consecutive deferrals	0.003	0.025	0.018	0.029	0.213
Days to previous donation	98 (84–167)	99 (90–182)	106 (77–168)	92 (70–147)	73 (59–118)
Hb in g/L	149 (142–157)	153 (147–159)	154 (147–162)	148 (142–156)	153 (142–163)
Proportion of Hb deferrals	0.004	0.022	0.018	0.029	0.129
First Hb level in g/L	150 (143–158)	154 (148–160)	155 (147–162)	150 (143–158)	153 (140–163)
Time of day as hour between 0 and 24	13.1 (10.8–15.6)	18.9 (17.8–19.7)	14.8 (13.1–16.4)	16.3 (13.1–18.7)	12.8 (11.2–14.6)
Hb level at previous visit in g/L	148 (139–156)	151 (143–158)	153 (144–161)	148 (140–155)	151 (137–162)
Proportion of low Hb at previous visit	0.003	0.020	0.018	0.030	0.124
Mean recent low Hb	0.008	0.066	0.074	0.127	0.553
Recent donations	4 (2–6)	4 (2–6)	5 (2–9)	5 (2–9)	4 (2–7)
Warm season proportion	0.500	0.477	0.491	0.494	0.524
Variable	Australia	Belgium	Women Finland	Netherlands	South Africa
Number of donors	10 000	8552	10 000	10 000	10 000
Age in years	41 (29–54)	39 (25–52)	53 (41–60)	52 (39–60)	44 (33–54)
Mean consecutive deferrals	0.003	0.025	0.018	0.029	0.213
Days to previous donation	98 (84–167)	99 (90–182)	106 (77–168)	92 (70–147)	73 (59–118)
Hb in g/L	149 (142–157)	153 (147–159)	154 (147–162)	148 (142–156)	153 (142–163)
Proportion of Hb deferrals	0.004	0.022	0.018	0.029	0.129
First Hb level in g/L	150 (143–158)	154 (148–160)	155 (147–162)	150 (143–158)	153 (140–163)
Time of day as hour between 0 and 24	13.1 (10.8–15.6)	18.9 (17.8–19.7)	14.8 (13.1–16.4)	16.3 (13.1–18.7)	12.8 (11.2–14.6)
Hb level at previous visit in g/L	148 (139–156)	151 (143–158)	153 (144–161)	148 (140–155)	151 (137–162)
Proportion of low Hb at previous visit	0.003	0.020	0.018	0.030	0.124
Mean recent low Hb	0.008	0.066	0.074	0.127	0.553
Recent donations	4 (2–6)	4 (2–6)	5 (2–9)	5 (2–9)	4 (2–7)
Warm season proportion	0.500	0.477	0.491	0.494	0.524

Table 8.1: Distributions of predictor variables in all five datasets. Numerical variables are described by their median and (interquartile range) unless otherwise stated. Dichotomous variables are described by the proportion of visits where the value was true.

Results

Table 8.1 shows the distribution of the predictor variables in all countries.

Hemoglobin measurements and deferral policies

All participating countries use hemoglobin measurements to defer donors, but there are differences in how hemoglobin is measured and when donors are deferred. Table 8.2 shows a summary of hemoglobin deferral related policies per country.

Country	When and how is Hb measured?	When is the donor deferred?
Australia	Capillary skin-prick Hb measurement by hemoglobinometer before each donation. If the Hb is below the threshold, a venous sample is taken from the non-donation arm and Hb is measured using the hemoglobinometer at the donation site to confirm.	Hb levels below 120 g/L (women) or below 130 g/L (men) as well as donors with a 20 g/L drop in Hb level relative to their previous donation.
Belgium	Hematology analyser Hb measurement from venous sample after every successful donation. Capillary skin-prick Hb measurement before donation for new donors and for donors with a venous Hb below the eligibility threshold at the previous donation.	Hb level below 125 g/L (women) or below 135 g/L (men) at previous and current donation.
Finland	Capillary skin-prick Hb measurement point of care (POC) before each donation. If the Hb is below threshold, venous sample is taken and Hb measured by POC device at donation site. [116]	Hb level below 125 g/L (women) or below 135 g/L (men) as well as donors with a 20 g/L drop in Hb level relative to their previous donation.
The Netherlands	Capillary skin-prick Hb measurement before each donation. If a Hb level is below the threshold, the measurement is repeated (up to three times in total). The highest value is used for the deferral decision. Since late 2017, donors are also deferred for low ferritin levels.	Hb level below 125 g/L (women) or below 135 g/L (men).
South Africa	Capillary skin-prick Hb measurement before each donation.	Hb level below 120 g/L (women) or below 130 g/L (men). Before 2020, cut-off levels of 125 and 135 g/L were used.

Table 8.2: Hemoglobin measurement and donor deferral policies per country.

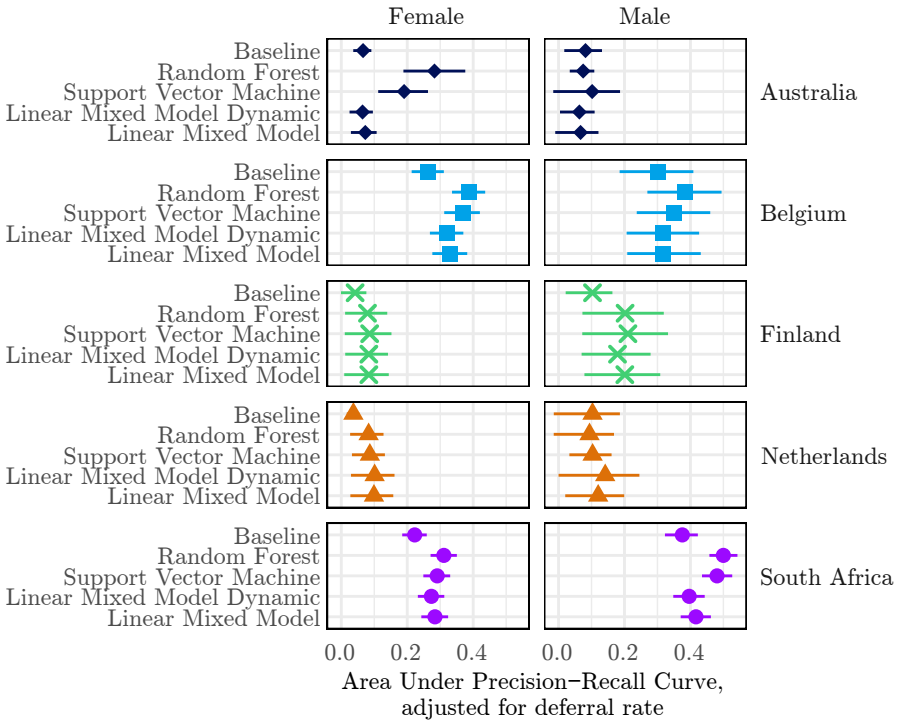

Figure 8.1: Area under the precision–recall (AUPR) curve for all countries and all models. Note that each AUPR curve is adjusted by subtraction of the country's deferral rate.

8

Comparison of model performance

Figure 8.1 shows the AUPR values (adjusted for deferral rate) and their confidence intervals for all models for all countries. All models outperform the baseline model in all countries. Performance of different models does not differ greatly within one country, except for Australian female donors, for which RF and SVM clearly outperform the LMM and DLMM. The same pattern is visible in South African male donors, although less obvious, and slightly in Belgium. In general, variation in within-country model performance is much smaller than variation between countries. Belgium and South Africa obtain significantly higher AUPR values than the other three countries in all models, except for the high-performing RF and SVM on Australian female donors.

Tables 8.3 and 8.4 show the predicted versus observed outcomes of the model with the lowest AUPR (baseline model, female donors, Finland; unadjusted AUPR = 0.07) and the model with the highest AUPR (RF, male donors, South Africa; unadjusted

	Observed outcome	
Predicted outcome	Accepted	Deferred
Accepted	1146	10
Deferred	807	37

Table 8.3: Observed versus predicted outcomes of the baseline model applied to female Finnish donors. This is the model with the lowest area under the precision–recall curve (0.07). The precision of class deferral is 0.04 and the recall is 0.79.

	Observed outcome	
Predicted outcome	Accepted	Deferred
Accepted	1433	108
Deferred	195	264

Table 8.4: Observed versus predicted outcomes of the baseline model applied to male South African donors. This is the model with the highest area under the precision–recall curve (0.69). The precision of class deferral is 0.58 and the recall is 0.71.

AUPR = 0.69) to illustrate the AUPRs with actual case counts to make the results more tangible.

Figure 8.2 shows the deferral rate and AUPR for all countries and models. Even though the AUPR values are adjusted for the deferral rate, there is still a positive correlation between deferral rate and (adjusted) AUPR. All models show the same pattern for this association. Again, we see that for Australian female donors the RF and SVM obtain a much higher AUPR than expected based on the deferral rate.

To further investigate whether the low deferral rates indeed affect the ability of the models to predict deferral, we intentionally modified the deferral rate of the Belgian datasets by removing a varying proportion of the deferred donors from the dataset and refitting the models on these adapted datasets. The results are shown in Figure 8.3. This figure clearly shows the positive association between deferral rate and AUPR. There is no monotonically increasing association even though the datasets with lower deferral rates are subsets of the datasets with larger deferral rates. The fact that classification tasks are more difficult when there is a large imbalance between outcome classes is a well-known phenomenon in statistics. [117]

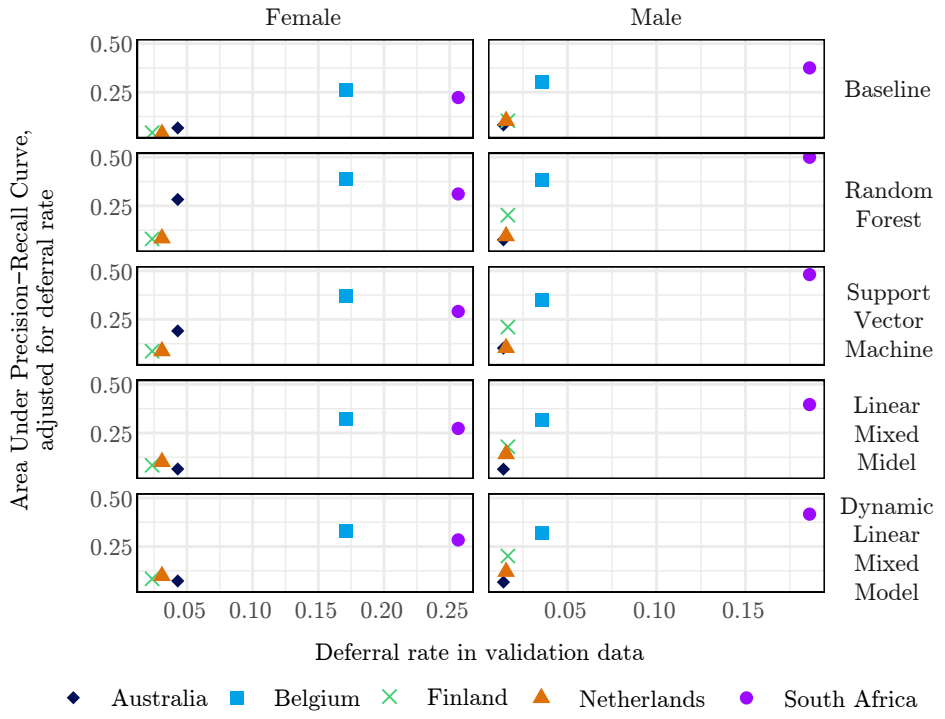

Figure 8.2: Adjusted area under the precision–recall value versus deferral rate in various settings for various models.

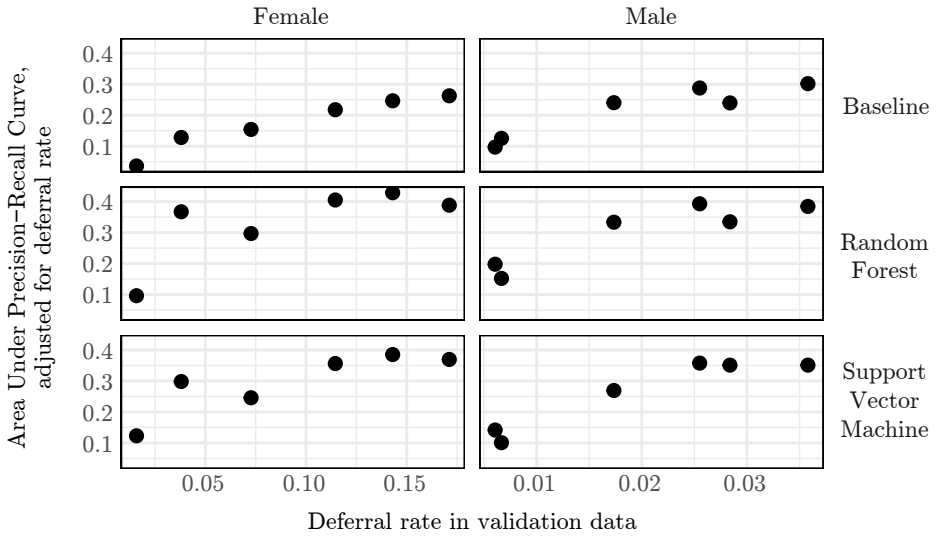

Figure 8.3: Adjusted area under the precision–recall as a function of the deferral rate for various deferral levels in the Belgian dataset. The reduction in deferral rate was obtained by sequentially removing an increasing number of deferred donations from the data.

Importance of individual variables

Figures 8.4 and 8.5 shows the variable importances derived from SHAP values calculated on a random subset of 1000 donors from the validation data. Variable importances are presented as mean absolute attribution (MAA) values. Variables are sorted by MAA over all countries and models (represented by the horizontal bars). For each individual country, the MAA values are provided and connected by a line.

RF and SVM

Comparing variable importances between countries within the same model allows identification of differences in predictive power of individual model parameters. In the RF and SVM models, previous hemoglobin is the most important predictor for all countries and sexes and has almost twice the MAA of the second-most important predictor. The MAA for most variables is similar across countries. There are some exceptions, however: for South Africa, the number of recent low hemoglobin measurements is much more important than in other countries, as well as the deferral status of the previous blood bank visit. For Belgium, whether the donation visit took place during

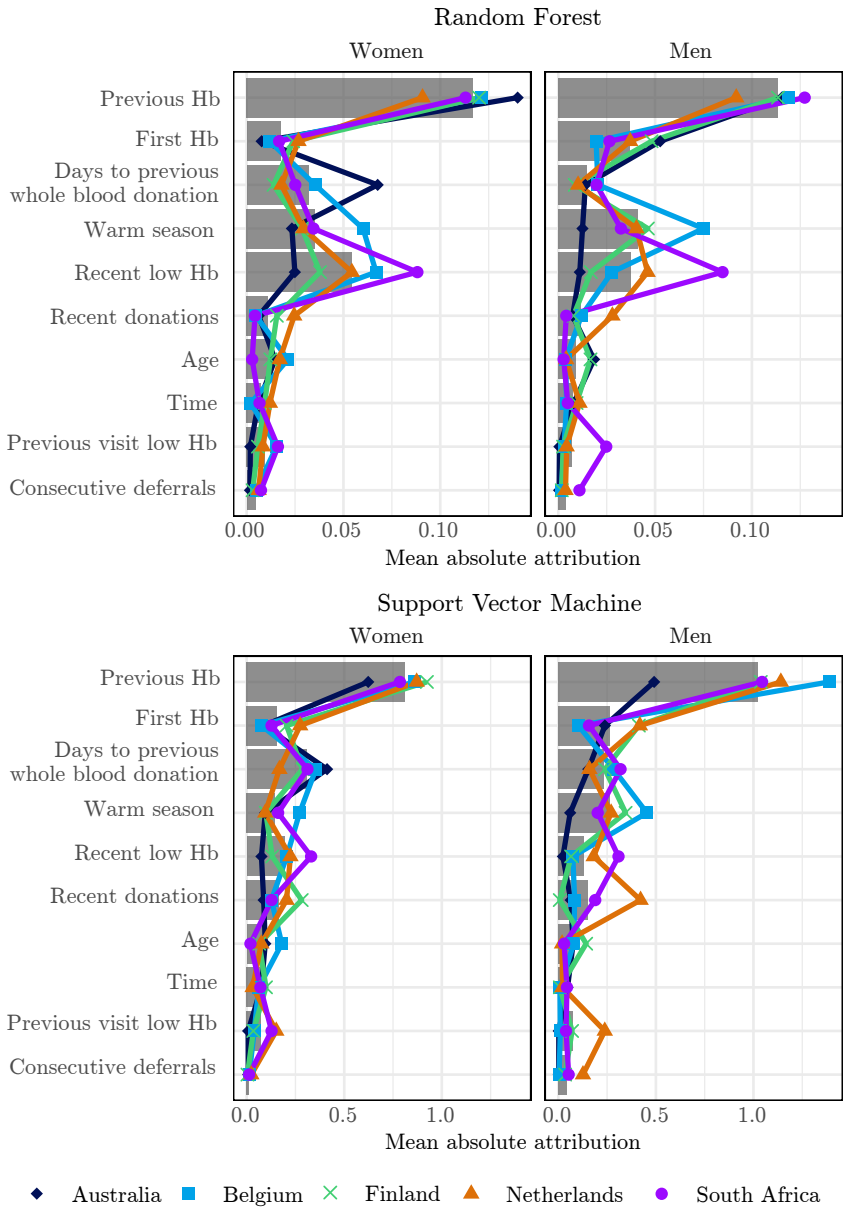

Figure 8.4: Variable importance (mean value and per individual country) determined by the mean absolute attribution according to SHapley Additive exPlanations values for the random forest and support vector machine models. The bars indicate the mean over all countries. Variables are ordered by the mean mean absolute attribution over both sexes.

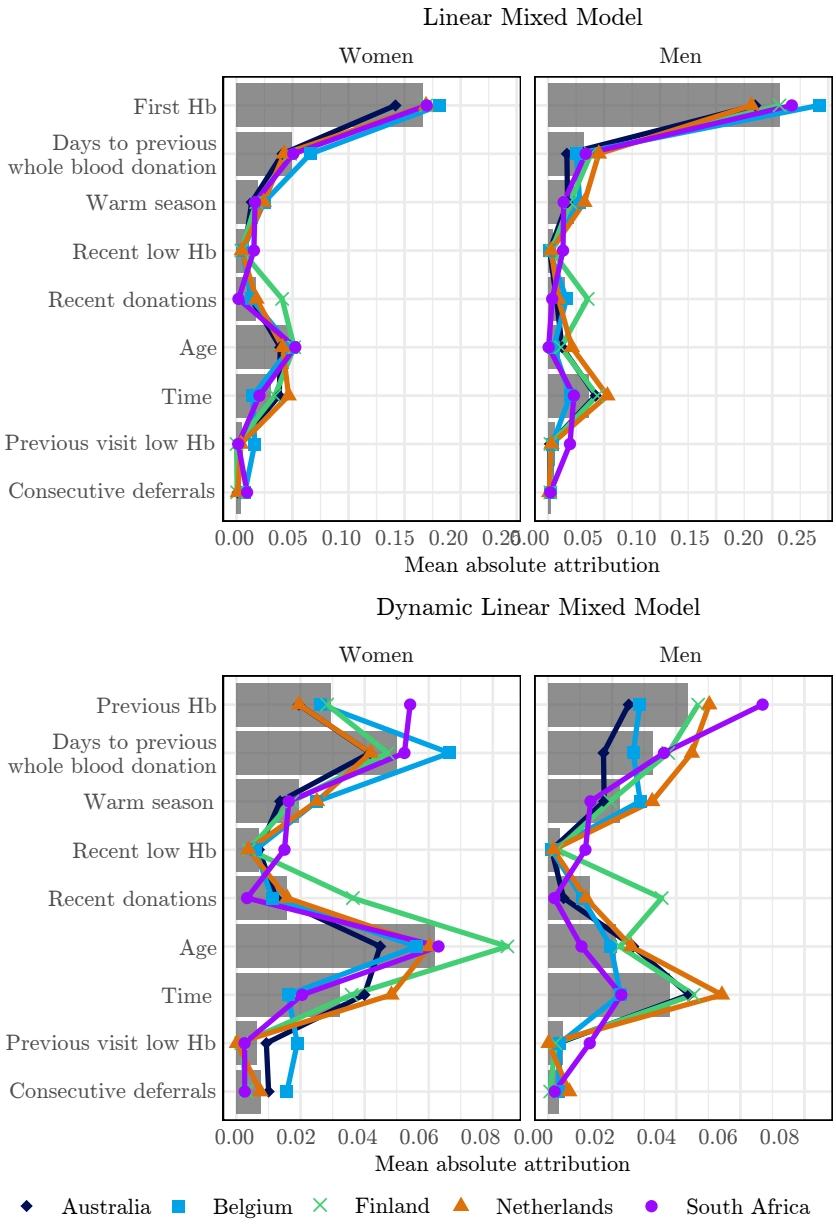

Figure 8.5: Variable importance (mean value and per individual country) determined by the mean absolute attribution according to SHapley Additive exPlanations values for the linear mixed model and dynamic linear mixed model. The bars indicate the mean over all countries. Variables are ordered by the mean mean absolute attribution over both sexes.

the warm season is more important than in the other countries.

Linear and dynamic linear mixed models

For the LMMs, the MAA of variables show the highest similarity between countries. A donor's first hemoglobin measurement is the most important predictor, and all other predictor variables have a relatively low MAA in comparison. Conversely, for DLMMs, there is much more variation in MAA values between countries and between sexes. For female donors, the most important predictor is age, and previous hemoglobin is only the third-most important predictor, which deviates considerably from what was found for all other models. In both LMM and DLLM, the difference in MAA for age between sexes is much larger than in RF and SVM models.

Unlike the RF and SVM models, the LMM and DLMM estimate regression coefficients that may be compared across countries. For consistency with other model results, we compared the MAA output rather than regression coefficients. A comparison of regression coefficients can be found in Supplementary Material. For all variables except for *Low Hb at previous visit* (which is the second to last most important predictor), coefficients are very similar between countries and 95% highest posterior density intervals mostly overlap.

Absolute value of MAA per model

It should be noted that the MAA values for different models are on different scales. In the baseline and SVM, SHAP values are on the log-odds scale, while for the RF and (dynamic) LMM, these are expressed on the probability scale. Since only the relative size of MAA values within models are compared, the difference in scales has no effect on the interpretation of the results.

The effect of sample size

We fitted the same models as above on the full datasets from Finland, the Netherlands and Australia to see whether this improves performance. This experiment showed that using the full dataset increases performance only by a very small amount and within the size of the confidence interval for the subsample of 10 000 donors.

Discussion

In this paper, various prediction models for hemoglobin deferral were applied to blood bank visit data from five countries to investigate the performance of prediction models in different settings. In all countries, the baseline was outperformed by all other models, although the overall performance was quite low for all models in all countries. Model performance, however, varies considerably between countries, and a high deferral rate is associated with better model performance. The relative importance of individual predictors is very similar in different countries. In particular, the hemoglobin level at previous donation is an important predictor for donor deferral in almost all models. This indicates that models learn the same associations in different settings, which supports the idea that these associations are the result of similar biological processes underlying donor deferral.

The similarity of the relative importance of predictors also indicates that the differences in performance are not caused by different associations between predictors and hemoglobin deferral. Rather, deferrals are more difficult to predict in countries with low deferral rates as there are fewer deferrals. The experiment with the Belgian data, which shows that the predictability collapses with a decrease in deferral rate, supports this finding. However, there appears to be an exception with the Australian data on female donors, where a relatively high AUPR is obtained for two models despite the very low deferral rate. Another possible explanation for the difference in performance could be that data collected in some countries is more informative than in others, for instance due to differences in the accuracy of hemoglobin measurements and/or differences in deferral policies. However, we were unable to confirm this as a plausible hypothesis: hemoglobin deferral is based on the same capillary measurement in South Africa and the Netherlands, and yet model performance on South African data is much higher than on Dutch data.

This study is the first to compare prediction models for hemoglobin deferral across different settings. By focusing on the comparison of models between countries rather than optimizing model performance based on variables available within a single country, the effect of the setting on model performance becomes visible. We show that low deferral rates substantially limit model performance, although they do not hinder the model in learning the same associations as with higher deferral rates. Comparing results for male donors from Australia and South Africa illustrates this perfectly: the deferral rate in South Africa is more than 10-fold than in Australia (18.6% vs. 1.4%), resulting in a much higher AUPR (0.50 vs. 0.08 for RF), yet the variable importance

is very similar.

Our findings are also in line with previously published work on hemoglobin deferral prediction, which consistently shows that previous hemoglobin measurements are by far the most important predictor. [99, 108, 110] Another interesting finding is that LMM, which is the only model to use a donor's first hemoglobin instead of the previous hemoglobin, performs just as well as the other models. This may indicate that most donors' hemoglobin levels are quite stable over time, and that predictions of personalised donation intervals can already be made after a first hemoglobin measurement at donor intake. To account for sudden drops in hemoglobin level, inclusion of the previous hemoglobin seems to be more relevant. The importance of first hemoglobin levels is also shown by others [118], which indicates that iron dynamics (hemoglobin and ferritin levels) in blood donors can be predicted over a longer period from the hemoglobin and ferritin levels at donor intake.

Although this study offers new insights into the predictability of donor deferral in different settings, the actual predictive value of the models is low, which may be explained by the substantial variability in hemoglobin measurement outcomes. [119] Note also that all analyses were done on donors with at least five donation attempts, which limits the generalisability of the models to the full donor population. Many blood banks collect more variables than were used in the predictions in this study and including those may improve model performance. Improved performance is paramount, as a model will create added value for the blood bank only when the benefits of the correctly predicted deferrals will outweigh the loss due to incorrectly predicted deferrals. The prediction of a potential reduction of donation intervals by some donors by the model may again add to the value of applying such prediction models.

Currently, the development of prediction models requires extensive expertise and data to enable prediction of donor deferral. Ideally, the work and insights developed by this collaboration would result in strategies that could also be of use to countries with limited resources.

In conclusion, this study shows that model architecture in most cases has a limited impact on the performance of prediction models for donor deferral, but in some cases, exemplified by Australia, certain model architectures can capture the data better than others. It would be recommended for any new country starting with hemoglobin deferral prediction to try several architectures if possible. Adding better predictor variables to the different model could considerably improve predictive performance. Performance is strongly affected by the donor deferral rate. For most countries with low deferral rates, prediction models are unlikely to contribute to an effective reduc-

tion of donor deferral rates. Conversely, deferral prediction models may be applied in countries with high deferral rates to reduce on-site deferral of donors. Hemoglobin deferral remains a relevant topic, as it negatively affects both donors and blood services. By joining efforts, we can enhance our understanding of which generic factors affect donor deferral and to what extent. Also, only by studying the performance in different settings, organization-specific and operational characteristics may be identified that enhance or deteriorate prediction models' performance, which may indicate directions for further research and meaningful policy changes.

Appendix

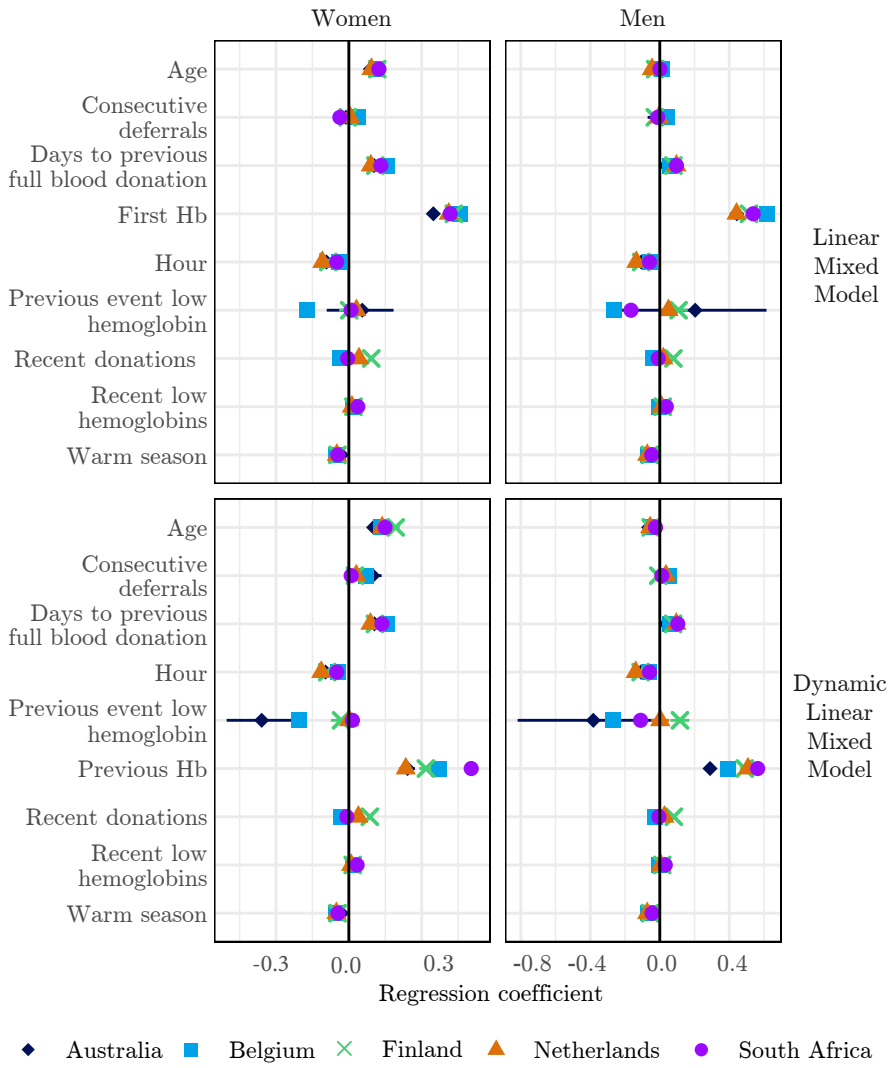

Figure S8.1: Regression coefficients per predictor for both linear models. The 95% highest posterior density intervals are indicated by horizontal lines (but not always visible due to being extremely narrow for many predictor variables).

	Linear Mixed Model - male donors				
	Australia	Belgium	Finland	Netherlands	South Africa
First Hb (g/L)	0.443	0.615	0.514	0.442	0.538
Days to previous whole blood donation	0.035	0.057	0.078	0.096	0.096
Warm season	-0.040	-0.065	-0.057	-0.072	-0.046
Recent low Hb	0.008	-0.006	0.017	0.010	0.037
Recent donations	-0.019	-0.039	0.079	0.020	-0.009
Age (years)	-0.030	0.012	-0.027	-0.045	0.001
Time (as hour between 0-24)	-0.106	-0.056	-0.107	-0.135	-0.060
Previous visit low Hb	0.204	-0.265	0.108	0.050	-0.166
Consecutive deferrals	-0.022	0.041	-0.029	-0.005	-0.012
	Linear Mixed Model - female donors				
	Australia	Belgium	Finland	Netherlands	South Africa
First Hb (g/L)	0.348	0.460	0.433	0.412	0.418
Days to previous whole blood donation	0.104	0.156	0.109	0.091	0.132
Warm season	-0.030	-0.053	-0.047	-0.050	-0.046
Recent low Hb	0.033	0.016	0.019	0.013	0.037
Recent donations	-0.035	-0.036	0.092	0.042	-0.004
Age (years)	0.089	0.113	0.117	0.093	0.124
Time (as hour between 0-24)	-0.092	-0.040	-0.085	-0.109	-0.050
Previous visit low Hb	0.055	-0.174	0.002	0.031	0.009
Consecutive deferrals	-0.012	0.037	-0.005	0.005	-0.037

Table S8.1: Regression coefficients per predictor for the Linear Mixed Models.

	Dynamic Linear Mixed Model - male donors				
	Australia	Belgium	Finland	Netherlands	South Africa
Previous Hb (g/L)	0.289	0.391	0.489	0.508	0.564
Days to previous whole blood donation	0.036	0.060	0.077	0.095	0.103
Warm season	-0.040	-0.067	-0.056	-0.073	-0.044
Recent low Hb	0.006	-0.006	0.014	0.006	0.031
Recent donations	-0.013	-0.027	0.081	0.027	-0.005
Age (years)	-0.064	-0.045	-0.052	-0.056	-0.024
Time (as hour between 0-24)	-0.109	-0.064	-0.110	-0.138	-0.059
Previous visit low Hb	-0.383	-0.268	0.116	0.003	-0.110
Consecutive deferrals	0.048	0.051	-0.009	0.036	0.012
	Dynamic Linear Mixed Model - female donors				
	Australia	Belgium	Finland	Netherlands	South Africa
Previous Hb (g/L)	0.242	0.369	0.318	0.234	0.504
Days to previous whole blood donation	0.105	0.157	0.108	0.089	0.137
Warm season	-0.030	-0.054	-0.046	-0.051	-0.045
Recent low Hb	0.030	0.016	0.016	0.010	0.034
Recent donations	-0.034	-0.031	0.086	0.039	-0.008
Age (years)	0.102	0.131	0.193	0.137	0.149
Time (as hour between 0-24)	-0.096	-0.045	-0.089	-0.114	-0.050
Previous visit low Hb	-0.359	-0.204	-0.032	0.001	0.015
Consecutive deferrals	0.099	0.072	0.025	0.031	0.010

Table S8.2: Regression coefficients per predictor for the Dynamic Linear Mixed Models.

CHAPTER

9

The added value of ferritin levels and genetic markers for the prediction of hemoglobin deferral

Published in: Vox Sanguinis 118(10): 825-834. doi:10.1111/vox.13517 Authors: M Vinkenoog, J Toivonen, M van Leeuwen, MP Janssen, M Arvas

Abstract

Background - On-site hemoglobin deferral for blood donors is sometimes necessary for donor health, but demotivating for donors and inefficient for the blood bank. Deferral rates could be reduced by accurately predicting donors' hemoglobin status before they visit the blood bank. Although such predictive models have been published, there is ample room for improvement in predictive performance. We aim to assess the added value of ferritin levels or genetic markers as predictor variables in hemoglobin deferral prediction models.

Methods - Support vector machines with and without this information (the full and reduced model, respectively) are compared in Finland and the Netherlands. Genetic markers are available in the Finnish data; ferritin levels in the Dutch data.

Results - While there is a clear association with hemoglobin deferral for both ferritin levels and several genetic markers, predictive performance increases only marginally with their inclusion as predictors. The recall of deferrals increases from 68.6% to 69.9% with genetic markers and from 79.7% to 80.0% with ferritin levels included. Subgroup analyses show that the added value of these predictors is higher in specific subgroups: e.g., for donors with minor alleles on SNP 17:58358769, recall of deferral increases from 73.3% to 93.3%.

Conclusions - Including ferritin levels or genetic markers in hemoglobin deferral prediction models improves predictive performance. The increase in overall performance is small, but may be substantial for specific subgroups. We recommend including this information as predictor variables when available, but not to collect it for this purpose only.

Introduction

Deferral of blood donors with low hemoglobin levels is necessary to prevent iron depletion. Currently, in Finland and the Netherlands, hemoglobin is measured before donation, and leads to on-site deferral if hemoglobin is below the donation threshold of 7.8 mmol/L (125 g/L) for women or 8.4 mmol/L (135 g/L) for men. On-site deferral is demotivating for donors and can be a reason to drop out of the donor pool permanently. [29] Hemoglobin deferral prediction models can help reduce the on-site deferral rate: for invitation-based donations, predictions can be included in the decision-making process of which donors to invite; for walk-in donations, the prediction could be communicated to the donor (e.g., shown on a donor dashboard or app that many blood banks offer), who can use this information to decide when to visit the blood bank.

Currently, hemoglobin deferral prediction models are not very accurate at predicting deferral on the specific day a donor may visit the blood bank. Although it is possible to correctly predict most deferrals as such (and therefore prevent them), this comes at the cost of incorrectly predicting some non-deferrals to be deferrals, which results in a large net loss of donations if these donors are then not invited to the blood bank based on this incorrect prediction. However, in a previous study we showed that predicting hemoglobin deferral at different time points, and inviting a donor once the predicted outcome is 'non-deferral', results in non-deferred donors to be invited earlier and deferred donors to be invited later, thereby eliminating the loss of successful donations. [109] This tells us that hemoglobin deferral prediction models are useful, and it is worth the effort of trying to improve the predictions.

Multiple studies [120, 110, 121] have shown previous hemoglobin levels to be the most important predictor of future hemoglobin deferral. Researchers from blood services in different countries have investigated many different potential predictors of hemoglobin deferral, to assess whether the inclusion of these predictors improves prediction performance. Most of these predictors were found to not substantially improve the models: information on menstruation, diet, ethnicity, and smoking all only slightly improve model performance, even though they are known to be associated with iron stores. [110] One small-scale study on 261 donors did show that ferritin, soluble transferrin receptor, and hepcidin were associated with subsequent anemia. [121]

In this study we investigate the added value of including ferritin levels and genetic information in hemoglobin deferral prediction models. Ferritin is routinely measured

at Sanquin, the Dutch national blood service, and therefore available for all donors. Genetic information for several iron-related SNPs is collected for many donors by the Finnish Red Cross blood service. Because the information in both countries is collected without targeting specific donors, our results provide a realistic indication of how much predictions would be improved if the prediction model was to be used in practice. Our results will therefore be useful for blood services that would like to collect additional donor information to improve hemoglobin deferral predictions.

Methods

Data

Data on blood donation attempts by whole-blood donors from (almost) five recent years were extracted from the eProgesa database (MAK-SYSTEM, Paris, France) in Finland and the Netherlands. Only data from donors who explicitly provided informed consent for the use of their data for scientific research were used. This consent is given by more than 99% of all Dutch donors. All Finnish blood donors studied provided an informed consent for biobank research in accordance with the Finnish Biobank Act and the study was approved by the Blood Service Biobank (project 004-2019). In Finland, approximately 23% of active blood donors have given this consent since the founding of the Blood Service Biobank in 2017.

Finnish data reflects data entries from January 2016 through April 2020, Dutch data from January 2017 through December 2021. For each visit the following information was collected in both countries: donor sex, donor age, donation date, and hemoglobin level. Additionally, ferritin level is measured at every new donor intake and upon every fifth donation in repeat donors in the Netherlands.

In Finland, only donors participating in the Blood Service Biobank are included, as only for these donors, genetic information related to iron metabolism is available. [122] The four SNPs were identified as significantly associated with higher prevalence of iron deficiency anemia in an iron deficiency anemia meta-analysis on Finnish and UK data. Polygenic risk scores were derived for three related endpoints: iron deficiency anemia, ferritin, and hemoglobin. [123]

In total, complete information on the predictor variables (see Table 9.1) was available for 172 508 donation attempts by 42 255 donors in Finland, and 456 384 donation attempts by 157 423 donors in the Netherlands.

The variable of interest is 'HbOK', a dichotomous variable that indicates whether

Variable	Unit or values	Description	Country/-ies
Sex	male, female	Biological sex of the donor; separate models are trained for men and women	Both
Age	years	Donor age at time of visit	Both
Month	1-12	Month of the year of the visit	Both
NumDon	donations	Number of successful (collected volume $>$ 250 ml) whole-blood donations in the last 24 months	Both
DaysSinceFirstDon	days	Number of days since the donor's first visit to the blood bank	Both
HbPrev$i$	mmol/L	Hemoglobin level at ith previous visit, for i between 1 and 5	Both
DaysSinceHbi	days	Time since related hemoglobin measurement at ith previous visit, for i between 1 and 5	Both
FerritinPrev	µg/L	Most recent ferritin level measured in this donor	Netherlands
SNP 1:169549811	0, 1, 2	Number of minor alleles in SNP rs6025	Finland
SNP 6:32617727	0, 1, 2	Number of minor alleles in SNP rs3129761	Finland
SNP 15:45095352	0, 1, 2	Number of minor alleles in SNP rs199138	Finland
SNP 17:58358769	0, 1, 2	Number of minor alleles in SNP rs199598395	Finland
PRS_anemia	standard deviations	Standardised polygenic risk score for anemia	Finland
PRS_ferritin	standard deviations	Standardised polygenic risk score for ferritin	Finland
PRS_hemoglobin	standard deviations	Standardised polygenic risk score for hemoglobin	Finland

Table 9.1: Predictor variables available in each country.

the result of the donation attempt was deferral (i.e., hemoglobin level below the eligibility threshold for donation) or non-deferral (i.e., hemoglobin level equal to or above the threshold).

Donor deferral due to low hemoglobin is similar in Finland and the Netherlands. Hemoglobin is measured using a capillary skin-prick device before each donation, and eligibility thresholds for donation are 7.8 mmol/L for women and 8.4 mmol/L for men. However, in case the measurement is below the eligibility threshold in Finland, hemoglobin is measured again (using the same device) in a venous sample, and this measurement is used for the deferral decision. In the Netherlands two additional capillary hemoglobin measurements are taken when the first measurement outcome is below the eligibility threshold, and the donor is allowed to donate if any of the three measurement outcomes is above the eligibility threshold.

Analyses

For both countries, two models were fitted for each sex: one with all predictor variables available (the full model), and one with only those predictor variables that are available in both countries (the reduced model). By comparing the full model with the reduced model in both countries, the added value of extra predictor variables (i.e., genetic information in Finland and ferritin information in the Netherlands) can be assessed. The prediction models used were based on models developed for an earlier study considering Dutch data only. [109] All models are based on support vector machines (SVMs), supervised machine learning models that learn a separation between outcome classes from a training set, after which the model can be used to predict donor deferral for observations in an unseen test set. Here the training set consists of blood bank visits in the first four years of data, whereas the test set consists of data collected in the final year.

Given a dataset and a set of predictor variables, a model consists of ten SVM sub-models. The sub-models are named SVM-sex-n, where sex indicates donor sex (m for male, f for female donors) and n indicates the number of previous blood bank visits that are used for prediction. That is, each sub-model includes HbPrev$i$ and DaysSinceHbi for i ranging from 1 to n as predictor variables. If sex is omitted in the sub-model name, it refers to the combination of two sex-specific sub-models. The number of blood bank visits (n) considered in this study varies from one through five, and so five sub-models per sex are created. Donors can only be included in the SVM-sex-n sub-model if they have at least n previous visits, therefore the sizes of the datasets used for both training and testing decrease from SVM-1 to SVM-5. Hyperparameters were

optimised separately for each sub-model, using stratified (on the outcome variable) five-fold cross-validation within the training set data only. Hyperparameters were optimised using grid search, using the balanced accuracy (defined as the weighted average of recall in both classes) as scoring method, which is suitable for datasets with imbalanced outcome sizes, as mistakes in the minority class are penalised more than those in the majority class.

During model training, the classification threshold is chosen again by optimizing the balanced accuracy. The predictive performance of the models is assessed using precision (also known as positive predictive value) and recall (also known as sensitivity) at this classification threshold. For non-deferral prediction, precision is defined as the proportion of true non-deferrals out of all predicted non-deferrals; recall is defined as the proportion of predicted non-deferrals out of all true non-deferrals. In this context, the complement of the precision is the hypothetical new deferral rate if the model would be used to choose which donors to invite, and the complement of the recall is the proportion of successful donations that would be missed by the model because the donors are incorrectly predicted to have a low hemoglobin level. Precision and recall can be calculated for both outcome classes ('deferral' and 'non-deferral').

The precision-recall curve is a graph in which the recall and the precision of a prediction model at varying classification thresholds is shown. The AUPR is the area under this curve, a number between 0 and 1, where 1 would indicate a perfect classifier. By subtracting the deferral rate from the AUPR, we get an adjusted AUPR, which reflects the improvement by the model over a strategy that would always predict non-deferral. Without this correction the improvement made by the model would be biased by the difference in deferral rate. The AUPR represents the ability of the model to distinguish between two classes at differing classification thresholds. It is possible for model A to have a higher AUPR than model B, even if precision and recall at the optimal classification threshold are the same in both models.

Model explanations

Because SVMs do not provide model coefficients that can be directly interpreted, we use Shapley Additive exPlanations (SHAP) values to investigate the importance of different predictor variables. [101] SHAP is a model agnostic explainer that shows the contribution of each predictor variable to the predicted outcome. This contribution is calculated for each individual observation separately (in a subsample of the test set) and is therefore very informative.

Subgroup analysis

To further investigate the value of including ferritin and genetic information in the models, we perform additional analyses in which donors are placed in groups defined by ferritin level or genotype. Deferral rate, model performance, and the difference between reduced and full model performance are calculated and compared to assess whether there are subgroups of donors for whom including the extra variables results in better predictions.

Software

All analyses were performed in Python 3.10 using packages numpy and pandas for data processing, scikit-learn for model training and predictions, shap for calculating SHAP values, and matplotlib for creating graphs. All code is available on GitHub and is indexed on Zenodo at https://doi.org/10.5281/zenodo.7780718.

Results

Table 9.2 shows the number of donation attempts used for each model in both countries. Deferral counts and rates are given in brackets. Sample sizes are much larger in the Netherlands than in Finland. This is because the total number of blood donations is much higher in the Netherlands than in Finland due to a larger population (17.4 million versus 5.5 million in 2020); but also, because genetic information is available in Finland in only a subgroup of donors, whereas ferritin measurements are available for all Dutch donors.

Deferral rates are very similar in both countries, around 3% for women and 1% for men. The biggest difference in deferral rates is found in men with at least one previous hemoglobin measurement, where the deferral rate is 0.6 percentage points higher in Finland. In most cases deferral rates go down whenever more previous visits are included; this is most likely the result of self-selection, where donors with lower hemoglobin levels are less likely to return for subsequent donations than donors with higher hemoglobin levels. Surprisingly, for Dutch men this pattern seems to some extent to be reversed as their deferral rate goes up with an increasing number of donations.

Tables S9.1 and S9.2 in the Appendix show the marginal distribution of the predictor variables, combined for all sub-models. Donors in Finland are older than donors in the Netherlands (median age 46 vs 30 years in women, 52 vs 34 years in men)

Model	Women		Men	
	Finland	Netherlands	Finland	Netherlands
SVM-1	83 628 (3216; 3.9%)	236 994 (7724; 3.3%)	88 880 (1480; 1.7%)	219 390 (2411; 1.1%)
SVM-2	68 718 (2494; 3.6%)	166 640 (5875; 3.5%)	78 268 (1264; 1.6%)	179 465 (2114; 1.2%)
SVM-3	55 011 (1859; 3.4%)	123 171 (4370; 3.6%)	68 225 (1054; 1.5%)	150 396 (1889; 1.3%)
SVM-4	43 164 (1307; 3.0%)	93 868 (3149; 3.4%)	58 951 (896; 1.5%)	127 807 (1667; 1.4%)
SVM-5	33 179 (868; 2.6%)	72 165 (2112; 2.9%)	50 540 (749; 1.5%)	108 832 (1424; 1.3%)

Table 9.2: Number of blood bank visits available per model for both countries; number of deferrals and deferral rates are given in brackets.

and the number of donations in the past two years ('NumDon') is also higher, with a difference in median donations of 2 for both sexes. This difference can be explained by the sample composition: the Finnish dataset consists of participants of the Blood Service Biobank, who have given consent for medical research and are typically regular, committed blood donors. Genetic information is only available for these donors.

Hemoglobin levels are slightly higher in Finland for both sexes for all variables HbPrev*i*, by 0.1-0.3 mmol/L. The time between subsequent donation attempts (variables DaysSinceHbi) is slightly shorter for Finnish women than for Dutch women, but almost identical for men. This difference can be partly explained by a difference in minimum donation interval between blood donations: for women, 91 days in Finland vs 122 days in the Netherlands; for men, 61 days in Finland vs 57 days in the Netherlands.

Predictive performance

Predictive performance can be assessed for individual sub-models, or for all sub-models combined, by using the most complex sub-model possible to predict each outcome. When more previous blood bank visits are taken into consideration, more predictor variables are used, and we expect the performance of the sub-model to increase. Figure 9.1 shows that this is the case for both the full and reduced model in both countries. The adjusted AUPR increases from SVM-1 through SVM-5 almost everywhere. An exception is the AUPR for class deferral in SVM-m-5, where the reduced model for Finnish donors shows an unexpected drop in the adjusted AUPR. For male donors,

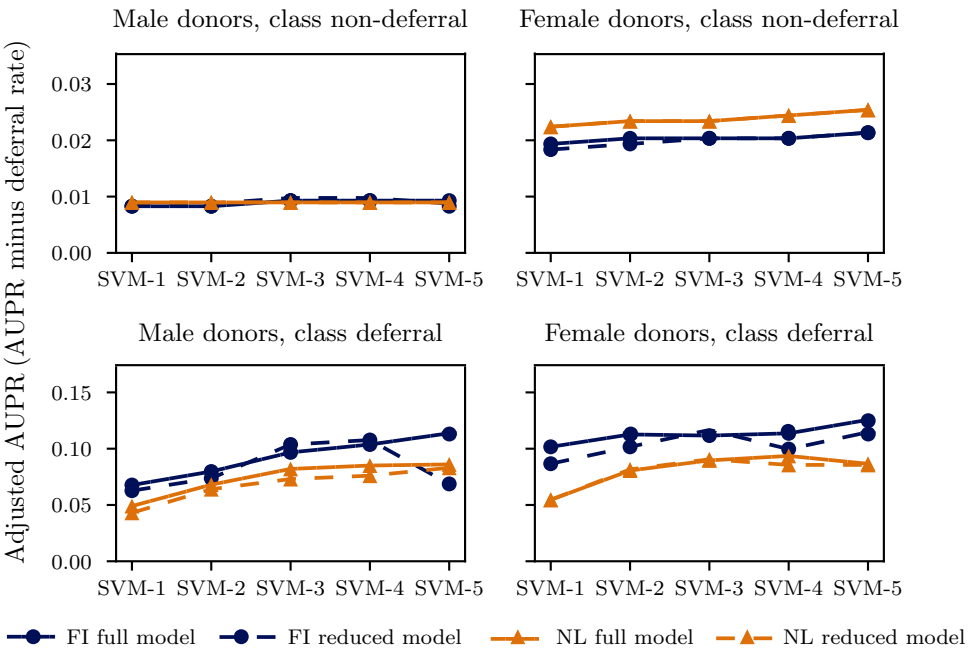

Figure 9.1: Adjusted AUPR by sub-model for both countries and both sets of predictor variables.

class non-deferral, the adjusted AUPR does not seem to change from SVM-m-1 through SVM-m-5.

Overall model performance and the difference in model performance between the full and reduced models are assessed by precision-recall curves and adjusted AUPR values as described in the Methods section. Figure 9.2 shows the precision-recall curves for various models (SVM-1 through SVM-5, using the model with the most predictor variables possible for each donation attempt) by sex and true outcome class. Table 9.3 shows the corresponding adjusted AUPR values for each model. In general, models are better at identifying non-deferrals (the most common outcome) than deferrals, even with scoring methods that weigh mistakes in both outcome classes proportionally. However, all curves are well above the baseline, indicating a structural improvement as compared to random guessing.

When comparing the reduced models to each other, one can observe that the performance is very similar in both countries. For women the AUPR is higher in Finland than in the Netherlands for the class deferral, but lower for the class non-

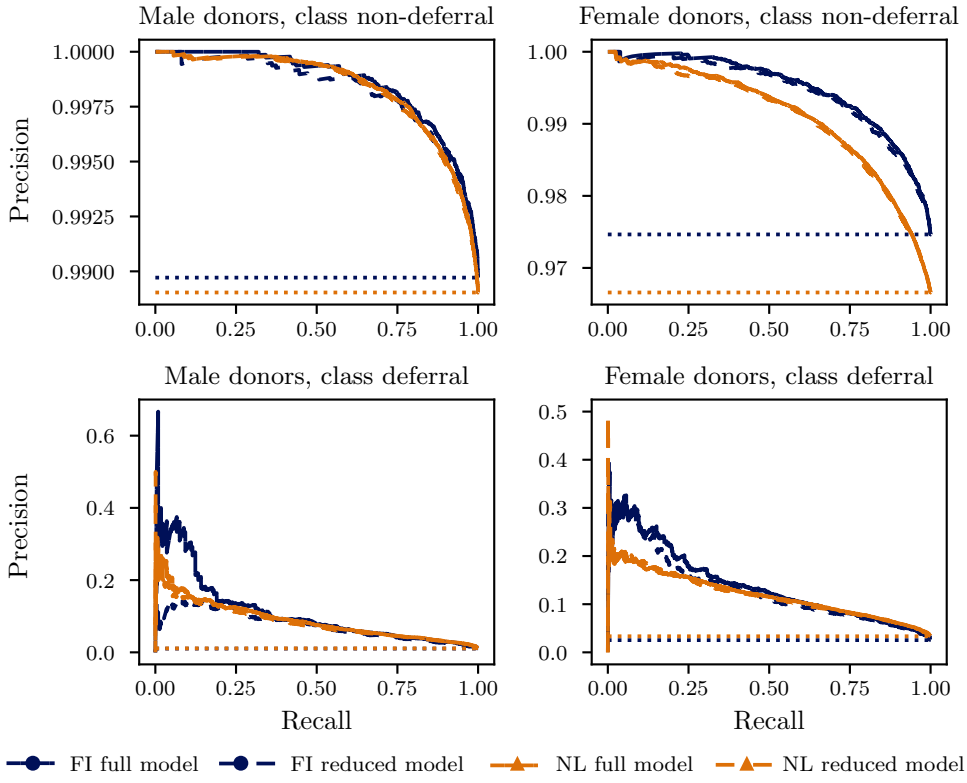

Figure 9.2: Precision-recall curves for the prediction models. For both countries, the curve is shown for the reduced and full prediction models. The baseline (proportion of observations belonging to this outcome class, i.e., for class deferral, the deferral rate) is shown as a dotted horizontal line.

	Baseline		Reduced model		Full model	
	FI	NL	FI	NL	FI	NL
Male donors, class non-deferral	0.990	0.989	0.008	0.009	0.009	0.009
Female donors, class non-deferral	0.975	0.967	0.019	0.024	0.020	0.024
Male donors, class deferral	0.010	0.011	0.066	0.072	0.104	0.078
Female donors, class deferral	0.025	0.033	0.106	0.086	0.115	0.086

Table 9.3: AUPR values for all models. AUPR values for the reduced and full models have been adjusted by subtracting the baseline AUPR.

deferral. This indicates that deferrals are more likely to be predicted correctly, but at the cost of more inaccuracies when predicting non-deferrals.

Moving from the reduced to the full model has virtually no effect on the AUPR for the class non-deferral: the AUPR of the full model is almost identical to the AUPR of the reduced model for both countries and sexes. For the class deferral, however, there is a difference: in Finland, AUPR increases by 58% (from 0.066 to 0.104) for men and by 8.5% (from 0.106 to 0.115) for women. In the Netherlands, AUPR remains the same for women (0.086 for both) but increases by 8.3% (from 0.072 to 0.078) for men.

Table 9.4 provides the confusion matrices of model predictions by the reduced and full models for both countries. In the Finnish data, going from the reduced to the full model causes 7 (1.9%) more deferrals to be predicted correctly, while 59 (0.3%) more non-deferrals are predicted correctly. These improvements were all for female donors; at the chosen threshold values, no net changes in the confusion matrix were seen for male donors. In the Dutch data, 13 (0.3%) more deferrals, as well as 1473 (1.0%) more non-deferrals are predicted correctly by the full model as compared to the reduced model.

Note that the large increase in AUPR for Finnish male donors, class deferral, is not reflected in the confusion matrices. The PR-curve in Figure 9.2 shows that the AUPR increase is due to higher precision in the full model between a recall of 0 and 0.2. However, the optimal classification threshold that is used by the models corresponds to a recall of 0.7, at which point precision in the full model is exactly equal to precision in the reduced model.

Variable importance

For all sub-models, SHAP values show the importance of the different predictor variables on the predicted outcome. Figures 9.3 and 9.4 shows SHAP plots of sub-model SVM-5 of the full model, separately for both sexes and countries. These plots show that in both countries and for both sexes, the most important predictor variable is HbPrev1, the most recent hemoglobin measurement. The direction of the association between the impact on the model output and the feature value for all HbPrev$i$ variables is sensible: a lower hemoglobin measurement is predictive of deferral. Age is a more important predictor variable for women than for men in both countries, which is known from previous studies: young women have the highest probability of being deferred due to low hemoglobin, due to monthly iron loss with menstruation.

The additional genetic and ferritin variables for either country end up rather low in the variable importance ranking. The importance of all polygenic risk score and

Finnish donors - reduced model		
	Predicted deferral	Predicted non-deferral
True deferral	363	166
True non-deferral	4573	18 713

Finnish donors - full model		
	Predicted deferral	Predicted non-deferral
True deferral	370 (+7)	159 (-7)
True non-deferral	4662 (-59)	18 624 (+59)

Dutch donors - reduced model		
	Predicted deferral	Predicted non-deferral
True deferral	3762	957
True non-deferral	56 676	145 549

Dutch donors - full model		
	Predicted deferral	Predicted non-deferral
True deferral	3775 (+13)	944 (-13)
True non-deferral	55 203 (-1473)	147 022 (+1473)

Table 9.4: Confusion matrices of predictions by the reduced and full models. Numbers are summed over both sexes and over all sub-models SVM¬-1 through SVM-5. Observations that can be predicted with multiple sub-models are included the most complex sub-model.

SNP variables in the Finnish models is very low. However, having the minor allele present in either SNP 6:32617727, SNP 15:45095354 or SNP 17:58358769 impacts the model output negatively. This effect is more pronounced in male than female donors.

Subgroup analysis in Finnish data

To further investigate the effect of the SNPs on deferral prediction, model performance was calculated for groups of donors with the same value for one SNP at a time. Donors with value 1 and 2 are grouped together, as the proportion of donors with value 2 is extremely low, except for the SNP on chromosome 6.

Table 4 shows that for the SNPs on chromosomes 1, 6 and 17, deferral rates are higher amongst donors with one or two minor alleles than in donors with only major alleles. As these SNPs are selected because of their association with iron deficiency or anemia, this is to be expected. Additionally, precision and recall of class deferral are generally higher for donors with minor alleles than for those without, for both the reduced and full models. The SNP 17:58358769 shows this same trend, but the difference between donors with and without minor alleles is much larger. Precision in this subgroup is about twice as high as the overall precision in both the reduced and full model. The increase in recall between the full and reduced model (which changes from 0.733 to 0.933) is the highest of all subgroups.

An additional analysis on the distribution of hemoglobin measurement per donor showed that the higher deferral rate among donors with minor alleles on SNP 17:58358769 can be explained through a combination of a slightly lower average hemoglobin level and a slightly higher variance. This causes these donors to have a slightly higher deferral probability (median 32.6% for donors without minor alleles, median 36.6% for those with minor alleles). This difference was not observed for the other SNPs.

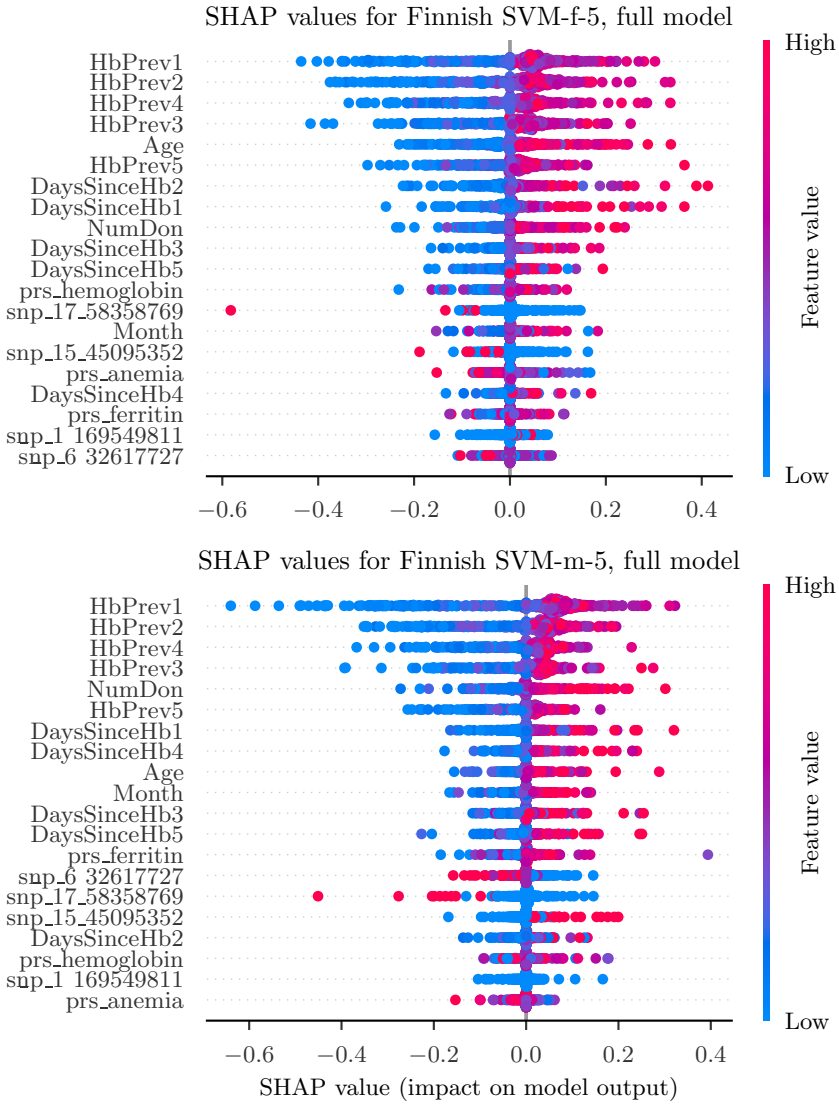

Figure 9.3: SHAP summary plots for the full Finnish model, for women (top) and men (bottom).

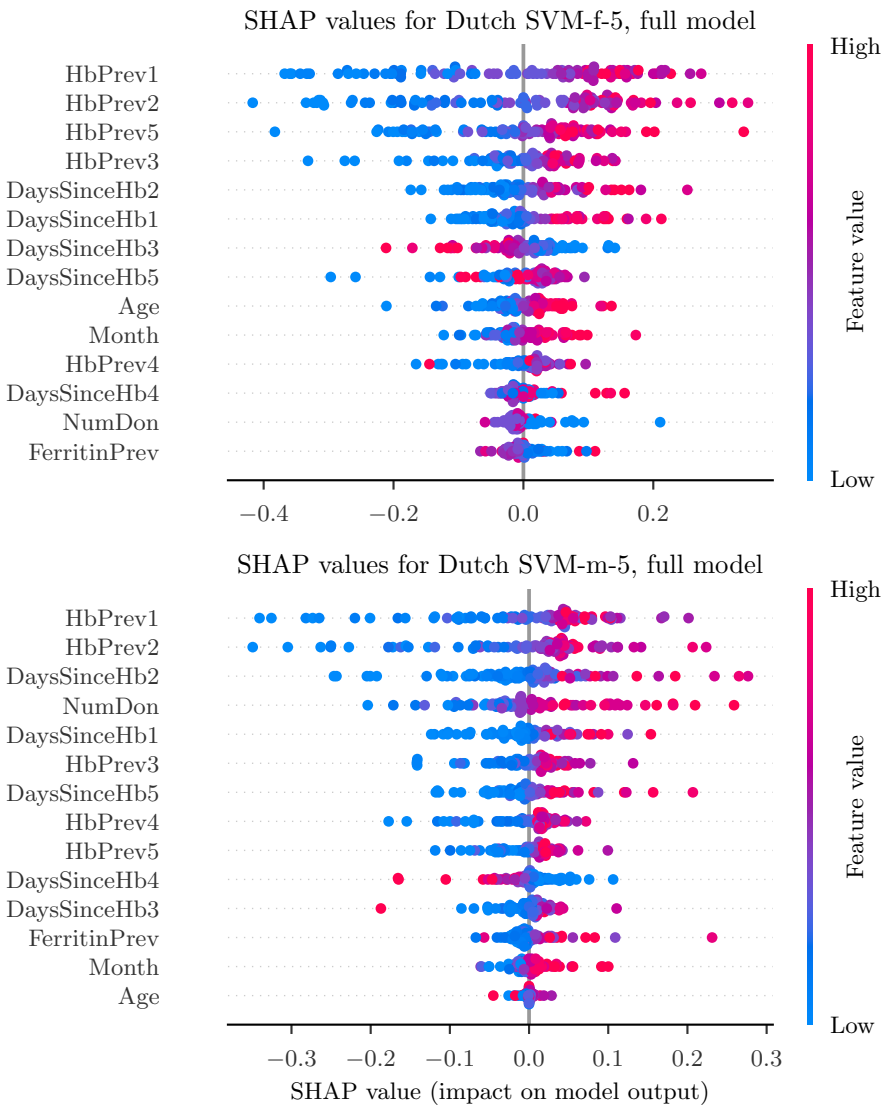

Figure 9.4: SHAP summary plots for the full Dutch model, for women (top) and men (bottom).

SNP	Minor alleles	N	Deferral rate	Precision (class deferral)		Recall (class deferral)	
				Reduced model	Full model	Reduced model	Full model
SNP 1:169549811	0	22 810	0.022	0.073	0.073	0.686	0.702
	1 or 2	1005	0.026	0.087	0.095	0.692	0.692
SNP 6:32617727	0	7268	0.021	0.063	0.067	0.573	0.587
	1	11 908	0.022	0.072	0.074	0.704	0.742
	2	4639	0.026	0.092	0.081	0.790	0.756
SNP 15:45095352	0	20 831	0.022	0.073	0.073	0.676	0.691
	1 or 2	2984	0.022	0.080	0.080	0.758	0.773
SNP 17:58358769	0	23 427	0.021	0.071	0.071	0.683	0.687
	1 or 2	388	0.077	0.156	0.129	0.733	0.933
Total	-	23 815	0.022	0.074	0.074	0.686	0.701

Table 9.5: Sample sizes, deferral rates, and precision and recall of outcome class deferral for subsets of donors based on values for four SNPs.

Subgroup analysis in Dutch data

Similar to the subgroup analysis in Finnish data, model performance was calculated for groups of donors with similar ferritin levels: $<$ 15 µg/L, 15-30 µg/L, 30-50 µg/L, 50-100 µg/L, and $>$ 100 µg/L. The first two groups are those that would be deferred for 12 or 6 months, respectively, in accordance with Sanquin's ferritin deferral policy.

Table 5 shows that precision and recall are highest for donors with ferritin levels between 30 and 50 µg/L. This is also the group of donors with the highest deferral rate: 3.2%, versus an overall deferral rate of 2.3%. The fact that this group has the highest deferral rate, and not donors with lower ferritin levels, can be explained by the fact that donors with ferritin levels below 30 µg/L were deferred for six months (twelve months for ferritin levels below 15 µg/L). This delay for the next donation provides the donors with sufficient time to replenish their iron stores and therefore reduces the deferral probability. Hence, donors with ferritin levels just above the ferritin-deferral threshold will have the highest hemoglobin-deferral rate, as they have neither the advantage of the donation break, nor that of a very high ferritin level, which also protects against low hemoglobin levels.

Ferritin level	N	Deferral rate	Precision (class deferral)		Recall (class deferral)	
			Reduced model	Full model	Reduced model	Full model
< 15 µg/L	7172	0.022	0.054	0.054	0.700	0.681
15 - 30 µg/L	19 903	0.022	0.058	0.056	0.744	0.783
30 – 50 µg/L	62 140	0.032	0.082	0.079	0.815	0.833
50 – 100 µg/L	65 141	0.024	0.064	0.063	0.798	0.799
> 100 µg/L	52 588	0.010	0.033	0.040	0.801	0.730
Total	206 944	0.023	0.062	0.064	0.797	0.800

Table 9.6: Sample sizes, deferral rates, and precision and recall of outcome class deferral for various subsets of donors based on their ferritin level.

Discussion

Predicting deferral for low hemoglobin levels is a topic of interest to many blood banks, as accurate predictions could aid in decreasing deferral rates. This study investigates the added value of including information on the donor's ferritin level or iron-related genetic information to improve hemoglobin deferral prediction. This is done by comparing prediction models with and without information on genetic markers and ferritin levels for the Finnish and Dutch blood bank respectively. The reduced models (i.e., without the additional information) use the exact same predictor variables in both countries. The increase in AUPR is larger for adding genetic markers than it is for adding ferritin levels. Especially for the Finnish male donors, including genetic markers in the prediction model improves the ability of the model to distinguish between the two outcome classes, although at the optimal classification threshold precision and recall do not increase from the reduced model. The SHAP values of the predictions by the full models in both countries show that both genetic markers and ferritin levels have a much smaller impact on the prediction than the variables included in the reduced models, as confirmed by the modest increase in AUPR between the reduced and full models.

Overall, including either genetic or ferritin information has little effect on the predictions made by the models. Both increase the proportion of deferrals that are predicted correctly: 1.9% and 0.3% more deferrals are correctly identified in the Finnish and Dutch setting respectively when the full model is used rather than the reduced model. However, we found that in both countries, there is a subgroup of donors for which the full model performs substantially better than the reduced model. These are Finnish donors with minor alleles on SNP 17:58358769, and Dutch donors with ferritin levels between 30-50 µg/L. In both cases, these are subgroups of donors with a higher-than-average deferral rate. Performance for these subgroups is already higher than average in the reduced model, but when using the full model this difference increases even further.

Other studies have shown that previous hemoglobin measurements are the most influential predictors for hemoglobin deferral. Including lifestyle behavior, smoking, ethnicity, or menstruation in prediction models also improves performance, but only marginally. [110] A Finnish study showed that genetic information does not improve the predictive performance of hemoglobin levels (as opposed to hemoglobin deferral). [108] This study confirms that the performance of prediction models increases slightly when either ferritin or genetic information is added. Still, considering the large number

of donation visits blood banks receive yearly, even a small increase could potentially prevent hundreds of deferrals. It should be noted that the Finnish population is more genetically homogenous than other countries, and that they are also genetically distinct from other countries due to several historic population bottlenecks and geographical isolation. [124] According to the Genome Aggregation Database (gnomAD) [125], the SNP 17:58358769 minor allele frequency in the Finnish population is 0.0147, and only 0.0007 in the European (non-Finnish) population. It is not found in any other populations and was discovered by an iron deficiency GWAS in the FinnGen project. [123] This means that findings on Finnish genetic data may not be representative for other countries, but analyses in other populations may discover similar population-specific variations that may make the use of genetic data more beneficial.

The main limitation of this study is that the effect of including ferritin and genetic information is studied in two different countries, rather than in a single population. By comparing against the reduced model and reporting the relative increase in performance, we attempt to mitigate this limitation. The very similar adjusted AUPRs of the reduced models and the similarity in SHAP values of the models indicate that the countries are rather comparable. A second limitation is that all Dutch donors could be included in this study, but only Finnish donors from the Blood Service Biobank, as genetic information is not available for other donors.

In general, we again confirm that accurately distinguishing deferrals from non-deferrals by predictive modelling is a complex task that comes at the cost of losing a substantial number of successful donations by incorrectly predicting them to be deferrals. A major reason for the low performance of our prediction models is the measurement variability, partly caused by the (pre-) analytical variability of the capillary hemoglobin measurements. [119] As long as we try to predict an outcome that is highly variable, the performance of any prediction model will remain unsatisfactory, regardless the number of predictor variables included.

However, in the absence of a better measurement or decision strategy, it is worthwhile investigating which information would lead to better hemoglobin deferral predictions as it still leads to a better understanding of the underlying process(es). Based on our results, we would recommend including ferritin and genetic information in prediction models in case these are readily available. Compared to the reduced model, including genetic information would have resulted in seven fewer deferrals and 59 more donations in one year, at a cost of genotyping approximately 24 000 donors. Including ferritin levels results in 13 fewer deferrals and 1473 more donations in one year, and although measuring ferritin levels is less expensive than genotyping, this measurement

must be repeated regularly whereas genotyping only has to be performed once for each donor. We would therefore not recommend collecting this information explicitly for the use in hemoglobin deferral prediction, as the marginal increase in performance is not likely to be worthwhile the investment of both time and money.

Appendix

	Women			
	Finland		Netherlands	
Number of donations	83 628		236 994	
Age	46	(29 - 57)	30	(23 - 47)
Month	6	(3 - 10)	7	(4 - 10)
NumDon	3	(2 - 5)	1	(0 - 3)
SNP_1_169549811	0: 79 991 1: 3567 2: 70		NA	
SNP_6_32617727	0: 26 241 1: 41 282 2: 16 105		NA	
SNP_15_45095352	0: 73 159 1: 10 101 2: 368		NA	
SNP_17_58358769	0: 82 336 1: 1287 2: 5		NA	
PRS_anemia ($*10^6$)	-0.002	(-0.847 - 0.828)	NA	
PRS_ferritin ($*10^6$)	0.032	(-1.191 - 1.280)	NA	
PRS_hemoglobin ($*10^6$)	0.039	(-3.010 - 3.105)	NA	
FerritinPrev	NA		47	(33 - 47)
HbPrev1	8.7	(8.3 - 9.1)	8.5	(8.1 - 8.9)
DaysSinceHb1	131	(104 - 203)	135	(104 - 194)
HbPrev2	8.7	(8.3 - 9.1)	8.5	(8.1 - 8.9)
DaysSinceHb2	280	(221 - 391)	301	(254 - 405)
HbPrev3	8.7	(8.3 - 9.1)	8.5	(8.1 - 8.8)
DaysSinceHb3	419	(338 - 558)	475	(396 - 627)
HbPrev4	8.7	(8.3 - 9.1)	8.4	(8.1 - 8.8)
DaysSinceHb4	546	(453 - 701)	653	(546 - 822)
HbPrev5	8.7	(8.3 - 9.1)	8.4	(8.1 - 8.8)
DaysSinceHb5	666	(561 - 825)	831	(703 - 1004)
Deferral rate	0.0385		0.0326	

Table S9.1: Marginal distribution of predictor variables in both countries for female donors. Variables are described by their median and 1st and 3rd quartiles, except for SNP variables, for which the allele count distributions are shown. Each donation attempt is included only once in this description and is given for the prediction using the highest number of previous visits only (e.g., a visit by a female donor with three previous visits could be included in SVM-f-1 through SVM-f-3 but is only included in SVM-f-3).

	Men			
	Finland		Netherlands	
Number of donations	88 880		219 390	
Age	52	(38 - 60)	34	(26 - 48)
Month	6	(3 - 10)	7	(4 - 10)
NumDon	5	(3 - 7)	3	(1 - 5)
SNP_1_169549811	0: 85 358 1: 3487 2: 35		NA	
SNP_6_32617727	0: 26 779 1: 43 714 2: 18 387		NA	
SNP_15_45095352	0: 78 223 1: 10 168 2: 489		NA	
SNP_17_58358769	0: 87 358 1: 1522 2: 0		NA	
PRS_anemia ($*10^6$)	-0.040	(-0.877 - 0.792)	NA	
PRS_ferritin ($*10^6$)	-0.023	(-1.272 - 1.243)	NA	
PRS_hemoglobin ($*10^6$)	-0.019	(-3.095 - 3.256)	NA	
FerritinPrev	NA		77	(44 - 141)
HbPrev1	9.6	(9.1 - 10.0)	9.4	(9.0 - 9.9)
DaysSinceHb1	98	(71 - 147)	81	(63 - 133)
HbPrev2	9.6	(9.1 - 10.0)	9.4	(9.0 - 9.8)
DaysSinceHb2	204	(154 - 293)	184	(138 - 287)
HbPrev3	9.6	(9.0 - 10.0)	9.4	(9.0 - 9.8)
DaysSinceHb3	306	(235 - 419)	300	(224 - 434)
HbPrev4	9.5	(9.0 - 10.0)	9.4	(8.9 - 9.8)
DaysSinceHb4	399	(314 - 535)	418	(314 - 581)
HbPrev5	9.5	(9.0 - 10.0)	9.4	(8.9 - 9.8)
DaysSinceHb5	489	(389 - 639)	535	(409 - 714)
Deferral rate	0.0167		0.0110	

Table S9.2: Marginal distribution of predictor variables in both countries for male donors. Variables are described by their median and 1st and 3rd quartiles, except for SNP variables, for which the allele count distributions are shown. Each donation attempt is included only once in this description and is given for the prediction using the highest number of previous visits only (e.g., a visit by a male donor with three previous visits could be included in SVM-m-1 through SVM-m-3 but is only included in SVM-m-3).

CHAPTER

10

Conclusions, general discussion and anticipated future research

Conclusions, general discussion and anticipated future research

Throughout this thesis, several statistical and data science analysis techniques have been applied to blood donation data in order to explore how their application can improve different aspects of donor management strategies. In this chapter, we summarise the findings from these analyses and discuss challenges that occurred in multiple chapters and therefore deserve more in-depth explanation and attention.

10.1 Conclusions

In the Introduction, seven research questions were formulated. Here, each research question is answered based on the results as described in Chapters 3-9, and main conclusions for each question are given.

10.1.1 Hemoglobin and ferritin levels

Q1 **Does a ferritin-based donor deferral policy prevent donors from returning with iron deficiency?**

The vast majority of donors that are deferred for low ferritin levels returns with considerably increased ferritin levels. After a 6-month deferral, ferritin levels were $\geq$ 15 µg/L for 88% of returning female donors and for 99% of returning male donors, which is a positive result. After a 12-month deferral, this was the case for 74% and 95% of returning female and male donors, respectively. [1] Although comparisons to a control group would be needed to draw conclusions about causality, it is reasonable to assume that if these donors had returned to the blood bank sooner (i.e., had they not been deferred), their ferritin levels would have been lower. From observational data, we also showed that the implementation of the ferritin deferral policy was associated with a substantial decrease in deferral rates due to low hemoglobin. Before the implementation of the policy, the hemoglobin deferral rate was around 8% for women and 5% for men, and currently it is down to 3% and 1%, respectively. [96]

In the same study, the distribution of ferritin levels was compared between sexes and between first-time and repeat donors. We found that in first-time donors, 25% of women and 2% of men have ferritin levels below the deferral threshold of 30 µg/L. These percentages are considerably higher in repeat donors, where 53%

of women and 42% of men have ferritin levels below 30 µg/L. The distribution of ferritin levels among first-time donors is very different for men and women, with women having much lower ferritin levels, but this difference almost disappears when comparing repeat donors. This suggests that regular blood donation results in a decrease in iron stores, which impacts men more than women because their natural iron stores are generally higher, and because of their higher donation frequency. These numbers underline the necessity of the ferritin-based deferral policy, especially since the proportion of new female donors under the age of 25 has been increasing rapidly in the Netherlands [34], and this is the group most at risk of having a low ferritin level.

The above findings are all described in Chapter 3. The main conclusion to this research question is that the ferritin-based donor deferral policy is successful in preventing donors from returning to donate with ferritin levels below 15 µg/L.

Q2 **What are determinants of variations in ferritin levels?**

Distributions of ferritin levels were compared between sexes and between first-time and repeat donors in the study on the effect of the ferritin-based deferral policy. [96] Within these groups, considerable variation in ferritin levels between individuals was observed. By applying structural equation modelling, we found that 25% of ferritin variance in new donors and 40% in repeat donors could be explained by individual characteristics, donation history (for repeat donors only), and environmental factors. [36] We confirmed previous findings, both our own and from other donor populations, that ferritin levels are substantially higher in men than in women among first-time donors, and that repeated blood donation impacts men's ferritin levels more than women's, resulting in similar ferritin levels for both sexes among repeat donors.

The main determinants of variation in ferritin levels are individual characteristics and donation history, as expected. The association between ferritin levels and environmental factors was smaller but still substantial. A likely explanation for this association is that air pollution can cause low-grade inflammation, and ferritin levels are known to be correlated with inflammatory activity. [47, 9, 126] Interestingly, the association was twice as high for repeat donors as for first-time donors. This indicates that environmental factors are more associated with ferritin recovery after blood loss than with ferritin levels in a steady state.

The above findings are described in Chapter 4. The main conclusion is that combining multiple determinants in a single integrative model allows us to ex-

plain a considerable part of ferritin variation based on individual characteristics, donation history and environmental factors.

Q3 **Can we find groups of donors whose hemoglobin levels change in a similar manner over the course of their donor career?**

By clustering donors' hemoglobin trajectories, we aimed to identify groups of donors with similar long-term responses to blood donation. This clustering could then be used to find associations with characteristics of the donors, and to find optimal donation intervals for the different groups. Both clustering methods resulted in distinct clusters of donors with clear differences in hemoglobin levels over time. However, the clusters differ mostly in the average hemoglobin value over time, as donors with similar hemoglobin levels at their first measurement are clustered together, and hemoglobin levels reduce slowly over time. With the clustering methods used, it was not possible to distinguish groups of donors with rapidly declining hemoglobin levels from those with relatively stable hemoglobin levels. [127] In later studies, the concept of clustering donors was replaced by making personalised predictions as described in research questions five through seven.

The results, along with an in-depth discussion on challenges in clustering these hemoglobin trajectories, are described in Chapter 5. The main conclusion is that the resulting clusters are based mostly on average hemoglobin value, therefore it seems more useful for the blood bank to focus on individual hemoglobin trajectories, rather than on characteristics that distinguish between clusters.

10.1.2 SARS-CoV-2 antibodies

Q4 **How are individual characteristics and symptoms associated with IgG antibody response in COVID-19 recovered donors?**

In this observational study into donors' IgG antibody response after a COVID-19 infection, we found higher age and BMI to be associated with higher antibody counts, indicating more severe illness. Antibody decay was found to be faster in male than in female donors, as well as for donors who had been hospitalised during their infection. We also identified associations between antibody counts and several self-reported symptoms that donors had experienced. The presence of nasal cold, headache and anosmia were associated with lower IgG levels, while dry cough, fatigue, fever, dyspnoea, diarrhoea, and muscle weakness were associated with higher IgG levels. [128] This was in line with findings from studies on

hospital cohorts, which found fatigue and dyspnoea to be prognostic for severe infection, while a stuffed nose (comparable to nasal cold in our data) was prognostic of mild infection. [91] At the time, our study was one of the largest studies concerning not only hospitalised patient cohorts, making it more representative of the total COVID-19 patient population.

These findings are described in Chapter 6. The main conclusion is that in addition to previously described associations with sex, age and BMI, SARS-CoV-2 antibody levels are also associated with several COVID-19 symptoms.

10.1.3 Prediction of hemoglobin deferral

Q5 **Can we accurately and reliably predict hemoglobin deferral based on historical data?**

We have presented a support vector machine (SVM) prediction model that predicts hemoglobin deferral based on several donor characteristics and up to five previous hemoglobin measurements. We found that although the model could correctly classify 80% of all deferrals, this comes at a cost of incorrectly classifying about 30% of donors with adequate hemoglobin levels as having to be deferred for low hemoglobin. This would imply a substantial net loss of donations for the blood bank. However, by using the model to predict deferral at different time points, we found that 64% of non-deferred donors would be invited earlier or on the same date, and 80% of deferred donors would be invited later. We assume that for some of these deferred donors, the extra recovery time would be enough to increase their hemoglobin level above the donation threshold. Using the prediction model to decide when to invite which donor, the deferral rate was estimated to decrease by 60% without decreasing the number of successful donations. [109]

SHAP values were used to see how predictor variables were related to the model prediction. These showed that previous hemoglobin levels are the most important predictors of future hemoglobin deferral, with low previous values being indicative of deferral. The use of SHAP values makes this model explainable, and we found that most predictor variables are related to the model predictions in ways that can be explained either by biological processes, or organisational policies.

These results are described in Chapter 6. The main conclusion is that using

prediction models to guide donor invitations may help to reduce donation intervals as well as deferral rates.

Q6 **How do country-specific blood bank policies and donor demographics affect hemoglobin deferral prediction models?**

By applying the same set of models to blood donation data of five different countries, we found that performance of the different models within countries is very similar. The most interesting result was that the relative importance of the predictor variables (again calculated using SHAP values) was very similar across countries. Previous hemoglobin remains undefeated as the best predictor variable for future hemoglobin deferral. Additionally, we found that model performance is highly dependent on the deferral rate, with higher deferral rates being associated with better model performance.

These findings are described in Chapter 8. The main conclusion is that model performance is more dependent on the deferral rate than on the model architecture, and that the relative importance of predictor variables is very similar across countries.

Q7 **Do ferritin measurements or genetic information add value to hemoglobin deferral prediction models?**

By comparing several simple models that only contain widely available predictor variables, we found that including ferritin as a predictor for hemoglobin deferral in Dutch donors increases model performance slightly, as does including genetic information as a predictor for Finnish donors. For certain subgroups of donors, including this extra information leads to a large increase in recall of deferrals. This is the case for donors with a rare minor allele on an iron-related single-nucleotide polymorphism (SNP) in Finland, and donors with ferritin levels just above the deferral threshold in the Netherlands.

The results are described in Chapter 9. The main conclusion is that although the overall value of ferritin and genetic information for hemoglobin deferral prediction is low, for specific subgroups it appears very useful in increasing the accuracy of deferral predictions.

10.2 General discussion

Throughout the research presented in this thesis, many challenges were encountered that were not study-specific and deserve a more in-depth discussion. Many studies were performed on the same dataset (albeit on updated versions), extracted from Sanquin's blood bank database system eProgesa. This dataset contains information on all donations that take place at Sanquin, and the purpose of recording this information is to ensure the safety and traceability of all blood products. During the pre-donation screening, donors are asked for their consent to the use of their data for scientific research, which over 99% of donors grant. Although many researchers at Sanquin use these data, it is not collected for research and therefore not optimised for that purpose.

Notable limitations of this dataset are the variability of recorded hemoglobin levels and the presence of selection bias. Two limitations for hemoglobin and ferritin research in blood donors in general are the uncertainty about how these proteins are related to health, and the fact that reproducing the research is difficult due to the very specific study population. These four topics are discussed in the following sections.

10.2.1 Hemoglobin measurement variability

Measurement variability is the phenomenon that whenever the same measurement is repeated the result will never be exactly the same. For measurements such as hemoglobin levels, measurement variability occurs through three causes. First, there is biological variation, as hemoglobin levels vary naturally throughout the day and year, and due to changes in diet, lifestyle, or even illness. Second, variation can occur as a result of differences in pre-analysis conditions, for instance by differences in temperature or transport, or how the donor physician handles the finger that the blood is collected from. Third and last, there is variation in the test itself, depending on the reliability of the measurement method, the assay and the machine used.

Because Sanquin tests hemoglobin at the new donor intake, and again approximately three weeks later before the first donation, we have data that allows estimation of the variability of the hemoglobin measurement. Differences due to biological variation in hemoglobin levels measured three weeks apart should be very minimal: diet and lifestyle are unlikely to change drastically in such a short time, and blood donation, pregnancy or major blood loss (apart from menstrual blood loss in premenopausal women) are also unlikely to have occurred. [11] From these two measurements, it can be derived that the standard deviation of an individual hemoglobin measurement is 0.43 mmol/L for men and 0.38 mmol/L for women. [119] This means that donors

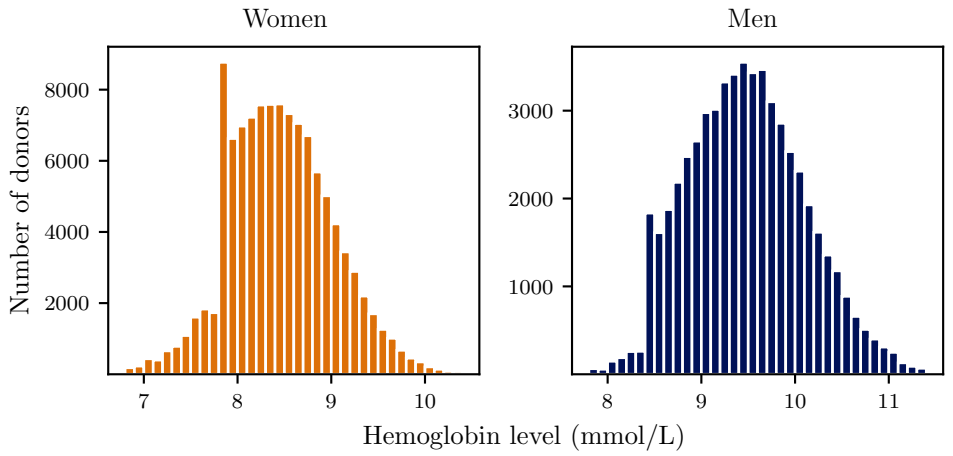

Figure 10.1: The distribution of recorded hemoglobin levels in all 114 459 female (left) and 58 511 male (right) prospective donors at donor intake between 2018 and 2020.

with hemoglobin levels around the deferral threshold will have a substantial chance of having a *measured* hemoglobin level below the threshold and therefore being deferred. The repeated measurement policy mitigates this effect, but at the same time introduces an upward bias in the data, as only the highest of three measurements is recorded. This bias is especially noticeable for donors around the donation threshold, and is illustrated by a simple histogram of all reported hemoglobin levels, as shown in Figure 10.1.

Clearly, the increase in observations between 7.7 and 7.8 mmol/L for women (and 8.3 and 8.4 mmol/L for men) is not due to those values naturally occurring more often, but rather it is an artefact of selective repeated measuring, and recording only the highest hemoglobin value. A more extensive review of this problem and several possible solutions was recently published and is well worth a read. [119]

10

The consequence of recording hemoglobin levels in this way is that a bias is introduced in the data, especially for hemoglobin levels around the deferral threshold. This makes the class imbalance (see Section 2.2.2) more extreme, making accurate classification harder. Additionally, observations right around the donation threshold potentially contain the most crucial information for deferral prediction, and precisely these observations are most impacted by the bias. It is therefore reasonable to expect that our models would perform better on data without such a bias.

10.2.2 Selection bias

Research on blood donors is generally influenced by the *healthy donor effect*, a selection bias caused by health criteria imposed on prospective blood donors. [129] The main consequence of the healthy donor effect is that it is difficult to draw conclusions on health effects of blood donation. Because donors are selected based on health criteria, in general they are ill less frequently than non-donors. This may lead to the incorrect conclusion that donating blood is beneficial for your health, while in reality, this association is found as a result of selection bias. [130] The effect persists during the entire donor career, as healthier donors are more likely to keep returning for subsequent donations.

Depending on the associations researched, it may therefore not be possible to generalise findings in donor cohorts to the general population. Most topics studies in this thesis would only be generalised to other donor populations, and therefore the healthy donor effect is not such a big complication. However, in some cases we may be tempted to extrapolate our findings to the general population, such as in the study on ferritin determinants, or the SARS-CoV-2 antibody paper. In the study on ferritin determinants, the association we found between environmental factors and ferritin level may have been underestimated, as this association is likely mediated by inflammation, and people with inflammation are probably underrepresented in the donor population. In the study on SARS-CoV-2 antibodies, only donors who were healthy enough to regularly donate plasma could be included, which means the results are mainly applicable towards people with a mild disease outcome. The results may not hold for people suffering from long COVID or other chronic health issues.

Due to this selection bias, it is also possible to draw incorrect conclusions or to miss correct ones when generalizing outside of donor cohorts. In some cases, the distribution of a predictor variable may be much narrower among donors than in the general population. This may prevent a true association from being found by a model, while in other cases, associations may be found that exist only in blood donors, as they are mediated by organisational policies of blood banks.

There is more selection bias in the donation dataset than the healthy donor effect. Ideally, we would like each donor to have the same probability of being invited to the blood bank and having their hemoglobin levels measured. Instead, donor invitations are based on many different criteria, the two most important being the current need for their blood group, and their response history. Donors that fail to respond to invitations (either by visiting the blood bank, or by rejecting the invitation) will be less prioritised and eventually may not at all be invited anymore. This is one example

of how the available data is based on existing donor management strategies, and we may not be able to learn an optimal strategy from such data.

In practice, this means that the performance of our hemoglobin deferral prediction models is inherently limited. Precision of deferral prediction is quite low, meaning that many donors with adequate hemoglobin levels are incorrectly predicted to have hemoglobin levels below the deferral threshold, which will lead to fewer donations and a potentially insufficient blood supply if these donors are not invited to the blood bank as a result of this incorrect prediction. However, predictions are only made for donors that were invited and visited the blood bank, and these donors are a (non-random) subset of all registered donors. If all donors were invited to the blood bank at the same rate, there would be a wider pool of donors for the model to choose from to mitigate the missed successful donations in absolute numbers. The fact that the process by which the donation dataset is formed is far from random also means that it is harder to predict what would happen if something were to be changed in the invitation process, such as the inclusion of a hemoglobin deferral prediction model. It is therefore difficult to predict what impact the application of such a model would have on the deferral rate exactly. In the current situation, a selection of loyal donors is prioritised for donor invitations, which leads to lower iron stores and higher probability of deferral due to low hemoglobin for these donors. Since our models are developed on data mostly from prioritised donors, it is possible that predictive performance on non-prioritised donors is lower.

10.2.3 Hemoglobin, ferritin and health

We monitor hemoglobin and ferritin levels in blood donors as an indication of their iron status, but research on the relation between these proteins and health is not entirely conclusive. Threshold values exist to diagnose anemia based on hemoglobin levels, but there are no clear threshold levels for hemoglobin and ferritin to diagnose iron deficiency without anemia. [9, 131] At Sanquin, donors are deferred for six months if their ferritin level is between 15 and 30 µg/L. This deferral is meant to prevent donors from returning with ferritin levels below 15 µg/L. However, among prospective female donors (women who have never donated blood before) 5% already have ferritin levels below 15 µg/L. [96] People generally only apply to become a blood donor when they feel healthy enough to do so, so these women are unlikely to experience symptoms of iron deficiency. On the other hand, the fact that they feel healthy enough to apply to become a blood donor does not exclude the possibility that they are already iron deficient or even anemic, as many women in the general population have a level of

iron deficiency, for example due to regular heavy menstrual blood loss, pregnancy, or breastfeeding. [132] This makes decisions on reference ranges of ferritin levels particularly challenging, and although those decisions are not in the scope of this thesis, it does complicate the interpretation of our study results to a wider health-related context. This is another reason that the results of our research are rarely generalisable outside of donation-related contexts, although it would be difficult to conceive of a relevant context where blood is regularly drawn without medical indication outside of blood banks.

Iron supplementation is often mentioned when discussing research concerning iron levels in blood donors. Would the best way to decrease deferral rates not be to provide donors with iron supplements? Some blood banks encourage all donors to take iron supplements, others encourage, or provide supplements to those most at risk for low iron (mainly young women, or donors donating at high frequencies). [133, 33] Sanquin does not recommend donors to take iron supplements, although of course donors are completely free to do so.

Even though iron supplementation is not current practice at Sanquin, its potential as a policy to enhance recovery after donating with low ferritin levels is currently being investigated. A randomised controlled trial is being conducted where donors with ferritin levels below 30 µg/L are given varying dosages of iron supplements or placebo pills. [134] Donor perceptions and changes therein are also important and being studied: as more donors are choosing to follow a vegetarian or vegan diet, their views on the necessity of iron supplements may also change. Furthermore, the success of iron supplementation policies is largely dependent on donors' willingness to take supplements, and their compliance.

10.2.4 Reproducibility of study results

Ideally, published research is reproducible by other research groups to be validated or challenged. Sanquin is the only blood bank in the Netherlands, making reproduction of our research by others difficult. Also, the data are considered privacy sensitive and therefore not easily shareable. It therefore makes sense to look across borders and compare our research results to those of other blood banks. The topics covered in this thesis are also studied in blood banks of other countries, and this allowed us to perform the comparison study presented in Chapter 8.

Collaboration with researchers from blood banks in Australia, Belgium, Finland, and South Africa showed that even though it appears that we do the same thing (hemoglobin testing and deferring donors below a certain cut-off value), small dif-

ferences in policies exist that make it hard to compare outcomes found in different countries. [20] Some sources of variation along with several (non-exhaustive) implementation alternatives are shown in Table 10.1, and an even wider range of alternative policies is described in a study by the BEST Collaborative Study Group. [55] All these factors are important to consider when comparing study results obtained in different settings, and it becomes even more complicated when we consider policy changes over the years. For instance, even only looking at Sanquin, ferritin testing was implemented in 2017, and hemoglobin deferral rates are now drastically lower than before 2017. What are the implications of this change in policy when comparing study outcomes from the Netherlands with those of other countries? Similarities found between countries may suggest similar associations, but any conclusions should be accompanied with words of caution for potential biases as a result of differences as specified in Table 10.1.

10.3 Anticipated developments and future research

Views on hemoglobin deferral and donor iron management are gradually changing. Researchers, policy-makers and health organisations are increasingly convinced that the most frequently used method of hemoglobin testing is suboptimal. One small change could be simply to record all hemoglobin measurements; even without changing the deferral policy, recording these extra measurements would allow obtaining unbiased hemoglobin estimates and better data for research (and decision-making in general). In general, it would be beneficial to move towards more individualised donation intervals rather than inviting donors back after a set amount of time and checking their hemoglobin and ferritin levels. Ferritin-guided donation intervals have been shown to increase ferritin and hemoglobin levels and thus decrease deferral rates. [135] These results are obtained with the same ferritin thresholds for each donor, but in the future donation intervals could even be guided individually for each donor, based on their own donation history.

The prediction models presented in this thesis are all strongly data-driven without any prior specification of how variables should theoretically be related to the outcome variable of interest. Currently, colleagues at Sanquin are obtaining great results with a prediction model based on ordinary differential equations with different states. [118] These states and equations are based on biological pathways for iron metabolism and erythropoiesis in the human body, and it turns out that this model captures changes in ferritin and hemoglobin very well. The current model requires only one

Source of variation	Implementation alternatives
Policy change	No policy change During (gradual) implementation of new policy After a change in policy
Timing of sampling for hemoglobin measurement	Before donation After donation
Method of sampling	Capillary, by finger-prick Venous sample, by venipuncture or via sampling bag
Repeated measurements	No repeated measurements Repeat same method if measurement is below threshold level Measure with different method if measurement is below threshold level
Hemoglobin deferral threshold	7.8 mmol/L for women, 8.4 mmol/L for men 7.4 mmol/L for women, 8.1 mmol/L for men
Additional iron and/or hemoglobin-related requirements	None Ferritin measurement Threshold for frop in hemoglobin relative to previous measurement
Maximum number of donations per year	Three for women, five for men One for women under 25, three for women over 25, five for men
Iron supplementation	Yes, provided or prescribed Yes, recommended No
Trigger to donate	Invitation-based Walk-in Mix of invitation-based and walk-in

Table 10.1: Sources of variation in donation policies and a non-exhaustive list of implementation differences between countries.

initial hemoglobin and ferritin measurement to predict subsequent levels and could potentially be improved by updating the model when new measurements are taken.

In our support vector machine models, we incorporated time by including up to five previous hemoglobin measurements as predictor variables, together with the time passed since these measurements. This way of including time-related variables is not optimal because the model does not allow linking the measurement results explicitly to the actual times when these measurements were obtained. It is therefore reasonable to expect that models that do incorporate such dependencies would perform better. Therefore, a potential other type of model worthwhile further exploration is the transformer network, which is a type of neural network architecture that has recently been used widely due to its high performance. [136] Although the most popular applications of transformer networks are in natural language processing (e.g., ChatGPT), they are very suitable for time series forecasting applications as well. [137] Transformer networks can model the relationship between measurements that are further apart, unlike recurrent neural networks. [136] However, a potential handicap may be the fact that the number of donation events from an individual donor are small relative to the length of the number of events for usual applications of transformer networks, which may limit the improvement in performance they may bring.

Hemoglobin deferral rates are currently very low in the Netherlands: about 3% and 1% of blood bank visits, by women and men respectively (reduced from 8% and 5% before implementation of the ferritin-based donor deferral policy). Any decrease in deferral rate is of course a good thing, and even a decrease of 0.5 percentage points would mean that 2000 on-site deferrals are prevented on a yearly basis. This potentially saves the blood bank the recruitment of several hundred new donors, as on-site deferral is known to be associated with donors dropping out of the donor pool. [29] However, although hemoglobin deferral rates currently are low, many donors are now deferred for low ferritin levels (approximately 10% of blood bank visits [135]), and therefore understanding how hemoglobin and ferritin are affected by blood donation remains extremely important.

The world of blood bank research has many opportunities for data science. More and more people see that data science can bring advances and improvements, but the actual implementation in day-to-day blood banking is far from easy. The primary task of blood banks will always be to ensure a safe and steady blood supply, and even a large increase in efficiency is not worth a small decrease in safety. Meanwhile, efforts are being made to make more room for data science research: Sanquin is preparing to set up the donor biobank *Sanquin Future Health*, which will process and store remainders

of blood donations to be used for research purposes. With repeated sampling and questionnaires from donors throughout the Netherlands, this biobank will be a treasure trove of data in a few years. As more data is collected, more complex models can be used to find new insights to enhance hemoglobin and ferritin predictions as well as inspire completely new research.

All blood banks struggle with the same balancing act: collecting enough donations to ensure a sufficient blood supply, while preventing iron deficiency and anemia in donors. As more insight is obtained in iron metabolism and how it is affected by blood donation, it will become clearer where there are opportunities to optimise donation strategies. Often, these are small steps that over the course of the coming years are likely to add up to substantial changes. In the future, this increased knowledge can lead to data-driven donation strategies, making optimal use of the information present in our data, resulting in a sufficient blood supply, maintained by healthy, motivated donors.

Bibliography

[1] C. Politis, J. C. Wiersum, C. Richardson, P. Robillard, J. Jorgensen, P. Renaudier, J.-C. Faber, and E. M. Wood. "The International Haemovigilance Network Database for the Surveillance of Adverse Reactions and Events in Donors and Recipients of Blood Components: technical issues and results". In: *Vox Sanguinis* 111.4 (2016), pp. 409–417.

[2] Ritchard G. Cable, Simone A. Glynn, Joseph E. Kiss, Alan E. Mast, Whitney R. Steele, Edward L. Murphy, David J. Wright, Ronald A. Sacher, Jerry L. Gottschall, Leslie H. Tobler, Toby L. Simon, and for the NHLBI Retrovirus Epidemiology Donor Study-II (REDS-II). "Iron deficiency in blood donors: the REDS-II Donor Iron Status Evaluation (RISE) study". In: *Transfusion* 52.4 (2012), pp. 702–711.

[3] Bryan R. Spencer and Alan E. Mast. "Iron status of blood donors". In: *Current Opinion in Hematology* 29.6 (Nov. 2022), pp. 310–316.

[4] Saurabh Zalpuri, Nienke Schotten, A Mireille Baart, Leo M van de Watering, Katja van den Hurk, and Marian GJ van Kraaij. "Iron deficiency–related symptoms in whole blood donors: a systematic review". In: *Transfusion* 59.10 (2019), pp. 3275–3287.

[5] Susan F. Clark. "Iron Deficiency Anemia". In: *Nutrition in Clinical Practice* 23.2 (2008), pp. 128–141.

[6] Laura Dean. "Blood and the cells it contains". In: *Blood Groups and Red Cell Antigens*. National Center for Biotechnology Information (US), 2005.

[7] M. Domenica Cappellini and Irene Motta. "Anemia in Clinical Practice-Definition and Classification: Does Hemoglobin Change With Aging?" In: *Seminars in Hematology* 52.4 (Oct. 2015), pp. 261–269.

[8] Nazanin Abbaspour, Richard Hurrell, and Roya Kelishadi. "Review on iron and its importance for human health". In: *Journal of Research in Medical Sciences : The Official Journal of Isfahan University of Medical Sciences* 19.2 (Feb. 2014), pp. 164–174.

[9] Axel Dignass, Karima Farrag, Jürgen Stein, et al. "Limitations of serum ferritin in diagnosing iron deficiency in inflammatory conditions". In: *International journal of chronic diseases* 2018 (2018).

[10] Maria Nieves Garcia-Casal, Sant-Rayn Pasricha, Ricardo X. Martinez, Lucero Lopez-Perez, and Juan Pablo Peña-Rosas. "Serum or plasma ferritin concentration as an index of iron deficiency and overload". In: *Cochrane Database of Systematic Reviews* 5 (2021).

[11] World Health Organization. *Serum ferritin concentrations for the assessment of iron status and iron deficiency in populations*. Tech. rep. World Health Organization, 2011.

[12] Nienke Schotten, Pieternel C. M. Pasker-de Jong, Diego Moretti, Michael B. Zimmermann, Anneke J. Geurts-Moespot, Dorine W. Swinkels, and Marian G. J. van Kraaij. "The donation interval of 56 days requires extension to 180 days for whole blood donors to recover from changes in iron metabolism". In: *Blood* 128.17 (Oct. 2016), pp. 2185–2188.

[13] Inge-Lis Kanstrup and Björn Ekblom. "Blood volume and hemoglobin concentration as determinants of maximal aerobic power". In: *Medicine & Science in Sports & Exercise* 16.3 (June 1984), p. 256.

[14] Kristin L. Sainani. "Explanatory Versus Predictive Modeling". In: *PM&R* 6.9 (Sept. 2014), pp. 841–844.

[15] Takaya Saito and Marc Rehmsmeier. "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets". In: *PLOS ONE* 10.3 (Mar. 2015), e0118432.

[16] Clement A Finch, James D Cook, Robert F Labbe, and Maria Culala. "Effect of blood donation on iron stores as evaluated by serum ferritin". In: (1977).

[17] Sareen S Gropper and Jack L Smith. *Advanced nutrition and human metabolism*. Cengage Learning, 2012.

[18] Joseph E. Kiss and Ralph R. Vassallo. "How do we manage iron deficiency after blood donation?" In: *British Journal of Haematology* 181.5 (2018), pp. 590–603.

[19] Esa T. Soppi. "Iron deficiency without anemia – a clinical challenge". In: *Clinical Case Reports* 6.6 (Apr. 2018), pp. 1082–1086.

[20] Joseph E. Kiss, Donald Brambilla, Simone A. Glynn, Alan E. Mast, Bryan R. Spencer, Mars Stone, Steven H. Kleinman, Ritchard G. Cable, and Lung for the National Heart and Blood Institute (NHLBI) Recipient Epidemiology and Donor Evaluation Study–III (REDS-III). "Oral Iron Supplementation After Blood Donation: A Randomized Clinical Trial". In: *JAMA* 313.6 (Feb. 2015), pp. 575–583.

[21] Maike G. Sweegers, Marian G.J. van Kraaij, and Katja van den Hurk. "First do no harm: iron loss in whole blood donors". In: *ISBT Science Series* 15.1 (2020), pp. 110–117.

[22] *Sanquin Annual Report 2018*.

[23] N. Milman, M. Kirchhoff, and T. JØrgensen. "Iron status markers, serum ferritin and hemoglobin in 1359 Danish women in relation to menstruation, hormonal contraception, parity, and postmenopausal hormone treatment". In: *Annals of Hematology* 65.2 (Aug. 1992), pp. 96–102.

[24] Nils Milman and Marianne Kirchhoff. "Influence of blood donation on iron stores assessed by serum ferritin and haemoglobin in a population survey of 1433 Danish males". In: *European Journal of Haematology* 47.2 (1991), pp. 134–139.

[25] Emanuele Di Angelantonio et al. "Efficiency and safety of varying the frequency of whole blood donation (INTERVAL): a randomised trial of 45 000 donors". In: *The Lancet* 390.10110 (Nov. 2017), pp. 2360–2371.

[26] Romilla Mittal, Neelam Marwaha, Sabita Basu, Harsh Mohan, and A. Ravi Kumar. "Evaluation of iron stores in blood donors by serum ferritin". In: *Indian Journal of Medical Research* 124.6 (Dec. 2006), p. 641.

[27] Brian Custer, Karen S. Schlumpf, David Wright, Toby L. Simon, Susan Wilkinson, Paul M. Ness, and for the NHLBI Retrovirus Epidemiology Donor Study-II. "Donor return after temporary deferral". In: *Transfusion* 51.6 (2011), pp. 1188–1196.

[28] Brian Custer, Artina Chinn, Nora V Hirschler, Michael P Busch, and Edward L Murphy. "The consequences of temporary deferral on future whole blood donation". In: *Transfusion* 47.8 (2007), pp. 1514–1523.

[29] Marloes LC Spekman, Theo G van Tilburg, and Eva-Maria Merz. "Do deferred donors continue their donations? A large-scale register study on whole blood donor return in the Netherlands". In: *Transfusion* 59.12 (2019), pp. 3657–3665.

[30] Martin Falkingham, Asmaa Abdelhamid, Peter Curtis, Susan Fairweather-Tait, Louise Dye, and Lee Hooper. "The effects of oral iron supplementation on cognition in older children and adults: a systematic review and meta-analysis". In: *Nutrition Journal* 9.1 (Jan. 2010), p. 4.

[31] Jed B Gorlin. "Iron replacement: precautionary principle versus risk-based decision making". In: *Transfusion* 59.5 (May 2019), pp. 1613–1615.

[32] Andreas S. Rigas, Ole B. Pedersen, Cecilie J. Sørensen, Erik Sørensen, Sebastian R. Kotzé, Mikkel S. Petersen, Lise W. Thørner, Henrik Hjalgrim, Christian Erikstrup, and Henrik Ullum. "No association between iron status and self-reported health-related quality of life in 16,375 Danish blood donors: results from the Danish Blood Donor Study". In: *Transfusion* 55.7 (2015), pp. 1752–1756.

[33] Graham A. Smith, Sheila A. Fisher, Carolyn Doree, Emanuele Di Angelantonio, and David J. Roberts. "Oral or parenteral iron supplementation to reduce deferral, iron deficiency and/or anaemia in blood donors". In: *Cochrane Database of Systematic Reviews* 7 (2014).

[34] Mindy Goldman, Whitney R. Steele, Emanuele Di Angelantonio, Katja van den Hurk, Ralph R. Vassallo, Marc Germain, Sheila F. O'Brien, and Biomedical Excellence for Safer Transfusion Collaborative (BEST) Investigators. "Comparison of donor and general population demographics over time: a BEST Collaborative group study". In: *Transfusion* 57.10 (2017), pp. 2469–2476.

[35] Sophie Waldvogel-Abramowski, Gérard Waeber, Christoph Gassner, Andreas Buser, Beat M. Frey, Bernard Favrat, and Jean-Daniel Tissot. "Physiology

of Iron Metabolism". In: *Transfusion Medicine and Hemotherapy* 41.3 (May 2014), pp. 213–221.

[36] Marieke Vinkenoog, Katja van den Hurk, Marian van Kraaij, Matthijs van Leeuwen, and Mart P Janssen. "First results of a ferritin-based blood donor deferral policy in the Netherlands". In: *Transfusion* 60.8 (2020), pp. 1785–1792.

[37] Stephen Kaptoge et al. "Longer-term efficiency and safety of increasing the frequency of whole blood donation (INTERVAL): extension study of a randomised trial of 20757 blood donors". In: *The Lancet Haematology* 6.10 (Oct. 2019), e510–e520.

[38] Bryan Spencer. "Blood donor iron status: are we bleeding them dry?" In: *Current opinion in hematology* 20.6 (Nov. 2013), pp. 533–539.

[39] A. Lecube, C. Hernández, D. Pelegrí, and R. Simó. "Factors accounting for high ferritin levels in obesity". In: *International Journal of Obesity* 32.11 (Nov. 2008), pp. 1665–1669.

[40] Tiffany C. Timmer, Rosa de Groot, Judith J.M. Rijnhart, Jeroen Lakerveld, Johannes Brug, Corine W.M. Perenboom, A. Mireille Baart, Femmeke J. Prinsze, Saurabh Zalpuri, C. Ellen van der Schoot, Wim L.A.M. de Kort, and Katja van den Hurk. "Dietary intake of heme iron is associated with ferritin and hemoglobin levels in Dutch blood donors: results from Donor InSight". In: *Haematologica* 105.10 (Nov. 2019), pp. 2400–2406.

[41] Susan J. Fairweather-Tait. "Iron nutrition in the UK: getting the balance right". In: *Proceedings of the Nutrition Society* 63.4 (Nov. 2004), pp. 519–528.

[42] Steven Bell, Andreas S. Rigas, Magnus K. Magnusson, Egil Ferkingstad, Elias Allara, Gyda Bjornsdottir, Anna Ramond, Erik Sørensen, Gisli H. Halldorsson, Dirk S. Paul, Kristoffer S. Burgdorf, Hannes P. Eggertsson, Joanna M. M. Howson, Lise W. Thørner, Snaedis Kristmundsdottir, William J. Astle, Christian Erikstrup, Jon K. Sigurdsson, Dragana Vuckovic, Khoa M. Dinh, Vinicius Tragante, Praveen Surendran, Ole B. Pedersen, Brynjar Vidarsson, Tao Jiang, Helene M. Paarup, Pall T. Onundarson, Parsa Akbari, Kaspar R. Nielsen, Sigrun H. Lund, Kristinn Juliusson, Magnus I. Magnusson, Michael L. Frigge, Asmundur Oddsson, Isleifur Olafsson, Stephen Kaptoge, Henrik Hjalgrim, Gudmundur Runarsson, Angela M. Wood, Ingileif Jonsdottir, Thomas F. Hansen, Olof Sigurdardottir, Hreinn Stefansson, David Rye, James E. Peters, David Westergaard, Hilma Holm, Nicole Soranzo, Karina Banasik, Gudmar Thorleifsson, Willem H. Ouwehand, Unnur Thorsteinsdottir, David J. Roberts, Patrick

Sulem, Adam S. Butterworth, Daniel F. Gudbjartsson, John Danesh, Søren Brunak, Emanuele Di Angelantonio, Henrik Ullum, and Kari Stefansson. "A genome-wide meta-analysis yields 46 new loci associating with biomarkers of iron homeostasis". In: *Communications Biology* 4.1 (Feb. 2021), pp. 1–14.

[43] Joseph E Kiss. "Laboratory and genetic assessment of iron deficiency in blood donors". In: *Clinics in laboratory medicine* 35.1 (2015), pp. 73–91.

[44] Bryan R. Spencer, Yuelong Guo, Ritchard G. Cable, Joseph E. Kiss, Michael P. Busch, Grier P. Page, Stacy M. Endres-Dighe, Steven Kleinman, Simone A. Glynn, Alan E. Mast, and For the National Heart, Lung, and Blood Institute Recipient Epidemiology and Donor Evaluation Study-III (REDS-III). "Iron status and risk factors for iron depletion in a racially/ethnically diverse blood donor population". In: *Transfusion* 59.10 (2019), pp. 3146–3156.

[45] Muriel Lobier, Johanna Castrén, Pia Niittymäki, Elina Palokangas, Jukka Partanen, and Mikko Arvas. "The effect of donation activity dwarfs the effect of lifestyle, diet and targeted iron supplementation on blood donor iron stores". In: *PLOS ONE* 14.8 (Aug. 2019), e0220862.

[46] Mohammed S Ellulu, Ismail Patimah, Huzwah Khaza'ai, Asmah Rahmat, and Yehia Abed. "Obesity and inflammation: the linking mechanism and the complications". In: *Archives of medical science: AMS* 13.4 (2017), pp. 851–863.

[47] Andrew J Ghio and Mitchell D Cohen. "Disruption of iron homeostasis as a mechanism of biologic effect by ambient air pollution particles". In: *Inhalation toxicology* 17.13 (2005), pp. 709–716.

[48] S. P. Doherty, C. Prophete, P. Maciejczyk, K. Salnikow, T. Gould, T. Larson, J. Koenig, P. Jaques, C. Sioutas, J. T. Zelikoff, M. Lippmann, and M. D. Cohen. "Detection of Changes in Alveolar Macrophage Iron Status Induced by Select PM2.5-Associated Components Using Iron-Response Protein Binding Activity". In: *Inhalation Toxicology* 19.6-7 (Jan. 2007), pp. 553–562.

[49] Rosa de Groot, Katja van den Hurk, Linda J. Schoonmade, Wim L. A. M. de Kort, Johannes Brug, and Jeroen Lakerveld. "Urban-rural differences in the association between blood lipids and characteristics of the built environment: a systematic review and meta-analysis". In: *BMJ Global Health* 4.1 (Jan. 2019), e001017.

[50] Jeroen Lakerveld, Alfred Wagtendonk, Ilonca Vaartjes, Derek Karssenberg, Jeroen Lakerveld, Brenda Penninx, Joline Beulens, Erik Timmermans, Martijn Huisman, Alfred Wagtendonk, Sophia Kramer, Marieke van Wier, Dorret Boomsma, Gonneke Willemsen, Carlo Schuengel, Mirjam Oosterman, Karien Stronks, Derek Karssenberg, Roel Vermeulen, Ilonca Vaartjes, Annemarie Koster, Coen Stehouwer, Katja van den Hurk, Eric Koomen, Renée de Mutsert, Margreet ten Have, Monique Verschuren, Susan Picavet, Mariëlle Beenackers, Frank van Lenthe, Arfan Ikram, Vincent Jaddoe, Tineke Oldehinkel, Trynke de Jong, Saakje Mulder, Aafje Dotinga, and GECCO Consortium. "Deep phenotyping meets big data: the Geoscience and hEalth Cohort COnsortium (GECCO) data to enable exposome studies in The Netherlands". In: *International Journal of Health Geographics* 19.1 (Nov. 2020), p. 49.

[51] James P Stevens. *Applied multivariate statistics for the social sciences*. Routledge, 2012.

[52] Cheng-Hsien Li. "Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares". In: *Behavior Research Methods* 48.3 (Sept. 2016), pp. 936–949.

[53] Li-tze Hu and Peter M. Bentler. "Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives". In: *Structural Equation Modeling: A Multidisciplinary Journal* 6.1 (Jan. 1999), pp. 1–55.

[54] Andreas Stribolt Rigas, Cecilie Juul Sørensen, Ole Birger Pedersen, Mikkel Steen Petersen, Lise Wegner Thørner, Sebastian Kotzé, Erik Sørensen, Karin Magnussen, Klaus Rostgaard, Christian Erikstrup, and Henrik Ullum. "Predictors of iron levels in 14,737 Danish blood donors: results from the Danish Blood Donor Study". In: *Transfusion* 54.3pt2 (2014), pp. 789–796.

[55] Saurabh Zalpuri, Bas Romeijn, Elias Allara, Mindy Goldman, Hany Kamel, Jed Gorlin, Ralph Vassallo, Yves Grégoire, Naoko Goto, Peter Flanagan, Joanna Speedy, Andreas Buser, Jose Mauro Kutner, Karin Magnussen, Johanna Castrén, Liz Culler, Harry Sussmann, Femmeke J. Prinsze, Kevin Belanger, Veerle Compernolle, Pierre Tiberghien, Jose Manuel Cardenas, Manish J. Gandhi, Kamille A. West, Cheuk-Kwong Lee, Sian James, Deanne Wells, Laurie J. Sutor, Silvano Wendel, Matthew Coleman, Axel Seltsam, Kimberly Roden, Whitney R. Steele, Milos Bohonek, Ramir Alcantara, Emanuele Di Angelantonio, Katja van den Hurk, and BEST Collaborative Study Group. "Variations in hemoglobin measurement and eligibility criteria across blood donation services are associated

with differing low-hemoglobin deferral rates: a BEST Collaborative study". In: *Transfusion* 60.3 (2020), pp. 544–552.

[56] Andrew J. Ghio, Joleen M. Soukup, Lisa A. Dailey, and Michael C. Madden. "Air pollutants disrupt iron homeostasis to impact oxidant generation, biological effects, and tissue injury". In: *Free Radical Biology and Medicine*. Air Pollution: Consequences for Cellular Redox Signaling, Antioxidant Defenses and Disease 151 (May 2020), pp. 38–55.

[57] Wenli Guo, Jie Zhang, Wenjun Li, Ming Xu, and Sijin Liu. "Disruption of iron homeostasis and resultant health effects upon exposure to various environmental pollutants: A critical review". In: *Journal of environmental sciences* 34 (2015), pp. 155–164.

[58] Jacob Westfall and Tal Yarkoni. "Statistically Controlling for Confounding Constructs Is Harder than You Think". In: *PLOS ONE* 11.3 (Mar. 2016), e0152719.

[59] Janet E Cade, Jennifer A Moreton, Beverley O'Hara, Darren C Greenwood, Juliette Moor, Victoria J Burley, Kairen Kukalizch, D Tim Bishop, and Mark Worwood. "Diet and genetic factors associated with iron status in middle-aged women". In: *The American Journal of Clinical Nutrition* 82.4 (Oct. 2005), pp. 813–820.

[60] H. A. Jackson, K. Carter, C. Darke, M. G. Guttridge, D. Ravine, R. D. Hutton, J. A. Napier, and M. Worwood. "HFE mutations, iron deficiency and overload in 10 500 blood donors". In: *British Journal of Haematology* 114.2 (2001), pp. 474–484.

[61] Erik Sørensen, Katrine Grau, Trine Berg, Anne Catrine Simonsen, Karin Magnussen, Christian Erikstrup, Morten Bagge Hansen, and Henrik Ullum. "A genetic risk factor for low serum ferritin levels in Danish blood donors". In: *Transfusion* 52.12 (2012), pp. 2585–2589.

[62] WW Hawkins, Eirlys Speck, and Verna G Leonard. "Variation of the hemoglobin level with age and sex". In: *Blood* 9.10 (1954), pp. 999–1007.

[63] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. "Time-series clustering – A decade review". In: *Information Systems* 53 (Oct. 2015), pp. 16–38.

[64] T Warren Liao. "Clustering of time series data—a survey". In: *Pattern recognition* 38.11 (2005), pp. 1857–1874.

[65] Sangeeta Rani and Geeta Sikka. "Recent techniques of clustering of time series data: a survey". In: *International Journal of Computer Applications* 52.15 (2012).

[66] Donald J Berndt and James Clifford. "Using dynamic time warping to find patterns in time series." In: *KDD workshop*. Vol. 10. Seattle, WA, USA: 1994, pp. 359–370.

[67] Stuart Lloyd. "Least squares quantization in PCM". In: *IEEE transactions on information theory* 28.2 (1982), pp. 129–137.

[68] Andreas Eckner. "Algorithms for unevenly-spaced time series: Moving averages and other rolling operators". In: *Working Paper*. 2012.

[69] Hasim Sak, Andrew W Senior, and Françoise Beaufays. "Long short-term memory recurrent neural network architectures for large scale acoustic modeling". In: (2014).

[70] Daniel Neil, Michael Pfeiffer, and Shih-Chii Liu. "Phased lstm: Accelerating recurrent network training for long or event-based sequences". In: *Advances in neural information processing systems* 29 (2016).

[71] Yu Zhu, Hao Li, Yikang Liao, Beidou Wang, Ziyu Guan, Haifeng Liu, and Deng Cai. "What to Do Next: Modeling User Behaviors by Time-LSTM." In: *IJCAI*. Vol. 17. 2017, pp. 3602–3608.

[72] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. "Recurrent neural networks for multivariate time series with missing values". In: *Scientific reports* 8.1 (2018), p. 6085.

[73] Nathan Post, Danielle Eddy, Catherine Huntley, May CI van Schalkwyk, Madhumita Shrotri, David Leeman, Samuel Rigby, Sarah V Williams, William H Bermingham, Paul Kellam, et al. "Antibody response to SARS-CoV-2 infection in humans: a systematic review (preprint)". In: (2020).

[74] Dianna L Ng, Gregory M Goldgof, Brian R Shy, Andrew G Levine, Joanna Balcerek, Sagar P Bapat, John Prostko, Mary Rodgers, Kelly Coller, Sandra Pearce, et al. "SARS-CoV-2 seroprevalence and neutralizing activity in donor and patient blood". In: *Nature communications* 11.1 (2020), p. 4698.

[75] Pyoeng Gyun Choe, Kye-Hyung Kim, Chang Kyung Kang, Hyeon Jeong Suh, EunKyo Kang, Sun Young Lee, Nam Joong Kim, Jongyoun Yi, Wan Beom Park, and Myoung-don Oh. "Antibody responses one year after mild SARS-CoV-2 infection". In: *Journal of Korean medical science* 36.21 (2021).

[76] Anu Haveri, Nina Ekström, Anna Solastie, Camilla Virta, Pamela Österlund, Elina Isosaari, Hanna Nohynek, Arto A Palmu, and Merit Melin. "Persistence of neutralizing antibodies a year after SARS-CoV-2 infection in humans". In: *European Journal of Immunology* 51.12 (2021), pp. 3202–3213.

[77] Maurice Steenhuis, Gerard van Mierlo, Ninotska IL Derksen, Pleuni Ooijevaar-de Heer, Simone Kruithof, Floris L Loeff, Lea C Berkhout, Federica Linty, Chantal Reusken, Johan Reimerink, et al. "Dynamics of antibodies to SARS-CoV-2 in convalescent plasma donors". In: *Clinical & translational immunology* 10.5 (2021), e1285.

[78] Erik H Vogelzang, Floris C Loeff, Ninotska IL Derksen, Simone Kruithof, Pleuni Ooijevaar-de Heer, Gerard van Mierlo, Federica Linty, Juk Yee Mok, Wim van Esch, Sanne de Bruin, et al. "Development of a SARS-CoV-2 total antibody assay and the dynamics of antibody response over time in hospitalized and non-hospitalized patients with COVID-19". In: *The Journal of Immunology* 205.12 (2020), pp. 3491–3499.

[79] Sabra L Klein, Andrew Pekosz, Han-Sol Park, Rebecca L Ursin, Janna R Shapiro, Sarah E Benner, Kirsten Littlefield, Swetha Kumar, Harnish Mukesh Naik, Michael J Betenbaugh, et al. "Sex, age, and hospitalization drive antibody responses in a COVID-19 convalescent plasma donor population". In: *The Journal of clinical investigation* 130.11 (2020), pp. 6141–6150.

[80] Mark Hamer, Catharine R Gale, Mika Kivimäki, and G David Batty. "Overweight, obesity, and risk of hospitalization for COVID-19: A community-based cohort study of adults in the United Kingdom". In: *Proceedings of the National Academy of Sciences* 117.35 (2020), pp. 21011–21013.

[81] Mohitosh Biswas, Shawonur Rahaman, Tapash Kumar Biswas, Zahirul Haque, and Baharudin Ibrahim. "Association of sex, age, and comorbidities with mortality in COVID-19 patients: a systematic review and meta-analysis". In: *Intervirology* 64.1 (2021), pp. 36–47.

[82] Nicholas S Hendren, James A De Lemos, Colby Ayers, Sandeep R Das, Anjali Rao, Spencer Carter, Anna Rosenblatt, Jason Walchok, Wally Omar, Rohan Khera, et al. "Association of body mass index and age with morbidity and mortality in patients hospitalized with COVID-19: results from the American Heart Association COVID-19 Cardiovascular Disease Registry". In: *Circulation* 143.2 (2021), pp. 135–144.

[83] Guillaume Plourde, Emanuel Fournier-Ross, Hubert Tessier-Grenier, Louis-Antoine Mullie, Michaël Chassé, and François Martin Carrier. "Association between obesity and hospital mortality in critical COVID-19: a retrospective cohort study". In: *International Journal of Obesity* 45.12 (2021), pp. 2617–2622.

[84] Jeffrey L Anderson, Heidi T May, Stacey Knight, Tami L Bair, Joseph B Muhlestein, Kirk U Knowlton, and Benjamin D Horne. "Association of sociodemographic factors and blood group type with risk of COVID-19 in a US population". In: *JAMA Network Open* 4.4 (2021), e217429–e217429.

[85] Mattia Miotto, Lorenzo Di Rienzo, Giorgio Gosti, Edoardo Milanetti, and Giancarlo Ruocco. "Does blood type affect the COVID-19 infection pattern?" In: *Plos one* 16.5 (2021), e0251535.

[86] Juanjuan Zhao, Quan Yuan, Haiyan Wang, Wei Liu, Xuejiao Liao, Yingying Su, Xin Wang, Jing Yuan, Tingdong Li, Jinxiu Li, et al. "Antibody responses to SARS-CoV-2 in patients with novel coronavirus disease 2019". In: *Clinical infectious diseases* 71.16 (2020), pp. 2027–2034.

[87] Maya F Amjadi, Sarah E O'Connell, Tammy Armbrust, Aisha M Mergaert, Sandeep R Narpala, Peter J Halfmann, S Janna Bashar, Christopher R Glover, Anna S Heffron, Alison Taylor, et al. "Specific COVID-19 symptoms correlate with high antibody levels against SARS-CoV-2". In: *Immunohorizons* 5.6 (2021), pp. 466–476.

[88] National Institute for Public Health and the Environment (RIVM). *Coronavirus Disease COVID-19*. Tech. rep.

[89] Seiya Yamayoshi, Atsuhiro Yasuhara, Mutsumi Ito, Osamu Akasaka, Morio Nakamura, Ichiro Nakachi, Michiko Koga, Keiko Mitamura, Kazuma Yagi, Kenji Maeda, et al. "Antibody titers against SARS-CoV-2 decline, but do not disappear for several months". In: *EClinicalMedicine* 32 (2021), p. 100734.

[90] Steven G Luke. "Evaluating significance in linear mixed-effects models in R". In: *Behavior research methods* 49 (2017), pp. 1494–1502.

[91] Jitian Li, Zhe Chen, Yifei Nie, Yan Ma, Qiaoyun Guo, and Xiaofeng Dai. "Identification of symptoms prognostic of COVID-19 severity: multivariate data analysis of a case series in Henan Province". In: *Journal of medical Internet research* 22.6 (2020), e19636.

[92] Catherine Gebhard, Vera Regitz-Zagrosek, Hannelore K Neuhauser, Rosemary Morgan, and Sabra L Klein. "Impact of sex and gender on COVID-19 outcomes in Europe". In: *Biology of sex differences* 11 (2020), pp. 1–13.

[93] Yifan Meng, Ping Wu, Wanrong Lu, Kui Liu, Ke Ma, Liang Huang, Jiaojiao Cai, Hong Zhang, Yu Qin, Haiying Sun, et al. "Sex-specific clinical characteristics and prognosis of coronavirus disease-19 infection in Wuhan, China: A retrospective study of 168 severe patients". In: *PLoS pathogens* 16.4 (2020), e1008520.

[94] Davide F Robbiani, Christian Gaebler, Frauke Muecksch, Julio CC Lorenzi, Zijun Wang, Alice Cho, Marianna Agudelo, Christopher O Barnes, Anna Gazumyan, Shlomo Finkin, et al. "Convergent antibody responses to SARS-CoV-2 in convalescent individuals". In: *Nature* 584.7821 (2020), pp. 437–442.

[95] Marloes LC Spekman, Steven Ramondt, and Maike G Sweegers. "Whole blood donor behavior and availability after deferral: consequences of a new ferritin monitoring policy". In: *Transfusion* 61.4 (2021), pp. 1112–1121.

[96] Marieke Vinkenoog, Katja van den Hurk, Marian van Kraaij, Matthijs van Leeuwen, and Mart P Janssen. "First results of a ferritin-based blood donor deferral policy in the Netherlands". In: *Transfusion* 60.8 (2020), pp. 1785–1792.

[97] Maike G. Sweegers, Saurabh Zalpuri, Franke A. Quee, Elisabeth M. J. Huis in 't Veld, Femmeke J. Prinsze, Emiel O. Hoogendijk, Jos W. R. Twisk, Anton W. M. van Weert, Wim L. A. M. de Kort, and Katja van den Hurk. "Ferritin measurement IN Donors—Effectiveness of iron Monitoring to diminish iron deficiency and low haemoglobin in whole blood donors (FIND'EM): study protocol for a stepped wedge cluster randomised trial". In: *Trials* 21.1 (Oct. 2020), p. 823.

[98] W Alton Russell, David Scheinker, and Brian Custer. "Individualized risk trajectories for iron-related adverse outcomes in repeat blood donors". In: *Transfusion* 62.1 (2022), pp. 116–124.

[99] AM Baart, WLAM De Kort, KGM Moons, and Y Vergouwe. "Prediction of low haemoglobin levels in whole blood donors". In: *Vox Sanguinis* 100.2 (2011), pp. 204–211.

[100] Kazem Nasserinejad, Joost van Rosmalen, Wim de Kort, Dimitris Rizopoulos, and Emmanuel Lesaffre. "Prediction of hemoglobin in blood donors using

a latent class mixed-effects transition model". In: *Statistics in medicine* 35.4 (2016), pp. 581–594.

[101] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017).

[102] William S Noble. "What is a support vector machine?" In: *Nature biotechnology* 24.12 (2006), pp. 1565–1567.

[103] Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. "Array programming with NumPy". In: *Nature* 585.7825 (2020), pp. 357–362.

[104] Wes McKinney et al. "Data structures for statistical computing in python". In: *Proceedings of the 9th Python in Science Conference*. Vol. 445. Austin, TX, 2010, pp. 51–56.

[105] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. "Scikit-learn: Machine learning in Python". In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.

[106] John D Hunter. "Matplotlib: A 2D graphics environment". In: *Computing in science & engineering* 9.03 (2007), pp. 90–95.

[107] Lauren Berkow. "Factors affecting hemoglobin measurement". In: *Journal of clinical monitoring and computing* 27 (2013), pp. 499–508.

[108] Jarkko Toivonen, Yrjö Koski, Esa Turkulainen, Femmeke Prinsze, Pietro della Briotta Parolo, Markus Heinonen, and Mikko Arvas. "Prediction and impact of personalized donation intervals". In: *Vox Sanguinis* 117.4 (2022), pp. 504–512.

[109] Marieke Vinkenoog, Matthijs van Leeuwen, and Mart P. Janssen. "Explainable haemoglobin deferral predictions using machine learning models: Interpretation and consequences for the blood supply". In: *Vox Sanguinis* 117.11 (2022), pp. 1262–1270.

[110] A. Mireille Baart, Tiffany Timmer, Wim L. A. M. de Kort, and Katja van den Hurk. "Lifestyle behaviours, ethnicity and menstruation have little added value in prediction models for low haemoglobin deferral in whole blood donors". In: *Transfusion Medicine* 30.1 (2020), pp. 16–22.

[111] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. "From local explanations to global understanding with explainable AI for trees". In: *Nature Machine Intelligence* 2.1 (Jan. 2020), pp. 56–67.

[112] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2020.

[113] Hadley Wickham, Romain François, Lionel Henry, Kirill Müller, and RStudio. *dplyr: A Grammar of Data Manipulation*. Sept. 2022.

[114] Hadley Wickham, Maximilian Girlich, and RStudio. *tidyr: Tidy Messy Data*. Sept. 2022.

[115] Hadley Wickham, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, Dewey Dunnington, and RStudio. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. May 2022.

[116] Sari Bäckman, Anne Valkeajärvi, Piia Korkalainen, Mikko Arvas, and Johanna Castrén. "Venous sample is superior to repeated skin-prick testing in blood donor haemoglobin second-line screening". In: *Vox Sanguinis* 115.8 (2020), pp. 617–623.

[117] J. A. ANDERSON. "Separate sample logistic discrimination". In: *Biometrika* 59.1 (Apr. 1972), pp. 19–35.

[118] Yared Paalvast, Sara Moazzen, Maike Sweegers, Boris Hogema, Mart Janssen, and Katja van den Hurk. "A computational model for prediction of ferritin and haemoglobin levels in blood donors". In: *British Journal of Haematology* 199.1 (2022), pp. 143–152.

[119] Mart P. Janssen. "Why the majority of on-site repeat donor deferrals are completely unwarranted..." In: *Transfusion* 62.10 (2022), pp. 2068–2075.

[120] Marieke Vinkenoog, Jarkko Toivonen, Tinus Brits, Dorien de Clippel, Veerle Compernolle, Surendra Karki, Amber Meulenbeld, Marijke Welvaert, Katja van den Hurk, Joost van Rosmalen, Emmanuel Lesaffre, Mikko Arvas, and Mart P Janssen. "An international comparison of hemoglobin deferral prediction models for blood banking". In: *Vox Sanguinis* (2023).

[121] Sant-Rayn Pasricha, Zoe K. McQuilten, Anthony J. Keller, and Erica M. Wood. "Hemoglobin and iron indices in nonanemic premenopausal blood donors predict future deferral from whole blood donation". In: *Transfusion* 51.12 (2011), pp. 2709–2713.

[122] Mitja I. Kurki et al. *FinnGen: Unique genetic insights from combining isolated population and national health register data*. Mar. 2022.

[123] Jarkko Toivonen, Johanna Castrén, FinnGen, and Mikko Arvas. "The Value of Genetic Data from 665,460 Individuals in Predicting Anemia and Suitability to Donate Blood". In: *Genetic Epidemiology* 46.7 (), pp. 477–477.

[124] Sanni Översti, Kerttu Majander, Elina Salmela, Kati Salo, Laura Arppe, Stanislav Belskiy, Heli Etu-Sihvola, Ville Laakso, Esa Mikkola, Saskia Pfrengle, Mikko Putkonen, Jussi-Pekka Taavitsainen, Katja Vuoristo, Anna Wessman, Antti Sajantila, Markku Oinonen, Wolfgang Haak, Verena J. Schuenemann, Johannes Krause, Jukka U. Palo, and Päivi Onkamo. "Human mitochondrial DNA lineages in Iron-Age Fennoscandia suggest incipient admixture and eastern introduction of farming-related maternal ancestry". In: *Scientific Reports* 9 (Nov. 2019), p. 16883.

[125] Konrad J. Karczewski, Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, Kristen M. Laricchia, Andrea Ganna, Daniel P. Birnbaum, Laura D. Gauthier, Harrison Brand, Matthew Solomonson, Nicholas A. Watts, Daniel Rhodes, Moriel Singer-Berk, Eleina M. England, Eleanor G. Seaby, Jack A. Kosmicki, Raymond K. Walters, Katherine Tashman, Yossi Farjoun, Eric Banks, Timothy Poterba, Arcturus Wang, Cotton Seed, Nicola Whiffin, Jessica X. Chong, Kaitlin E. Samocha, Emma Pierce-Hoffman, Zachary Zappala, Anne H. O'Donnell-Luria, Eric Vallabh Minikel, Ben Weisburd, Monkol Lek, James S. Ware, Christopher Vittal, Irina M. Armean, Louis Bergelson, Kristian Cibulskis, Kristen M. Connolly, Miguel Covarrubias, Stacey Donnelly, Steven Ferriera, Stacey Gabriel, Jeff Gentry, Namrata Gupta, Thibault Jeandet, Diane Kaplan, Christopher Llanwarne, Ruchi Munshi, Sam Novod, Nikelle Petrillo, David Roazen, Valentin Ruano-Rubio, Andrea Saltzman, Molly Schleicher, Jose Soto, Kathleen Tibbetts, Charlotte Tolonen, Gordon Wade, Michael E. Talkowski, Benjamin M. Neale, Mark J. Daly, and Daniel G. MacArthur. "The mutational constraint spectrum quantified from variation in 141,456 humans". In: *Nature* 581.7809 (May 2020), pp. 434–443.

[126] Kate F Kernan and Joseph A Carcillo. "Hyperferritinemia and inflammation". In: *International Immunology* 29.9 (Nov. 2017), pp. 401–409.

[127] Marieke Vinkenoog, Mart Janssen, and Matthijs van Leeuwen. "Challenges and limitations in clustering blood donor hemoglobin trajectories". In: *Advanced Analytics and Learning on Temporal Data: 4th ECML PKDD Workshop, AALTD 2019, Würzburg, Germany, September 20, 2019, Revised Selected Papers 4*. Springer International Publishing. 2020, pp. 72–84.

[128] Marieke Vinkenoog, Maurice Steenhuis, Anja ten Brinke, JG van Hasselt, Mart P Janssen, Matthijs van Leeuwen, Francis H Swaneveld, Hans Vrielink, Leo van de Watering, Franke Quee, et al. "Associations between symptoms, donor characteristics and IgG antibody response in 2082 COVID-19 convalescent plasma donors". In: *Frontiers in immunology* 13 (2022), p. 637.

[129] Femke Atsma, Ingrid Veldhuizen, André Verbeek, Wim de Kort, and Femmie de Vegt. "Healthy donor effect: its magnitude in health research among blood donors". In: *Transfusion* 51.8 (2011), pp. 1820–1828.

[130] Franke A. Quee, Karlijn Peffer, Anique D. Ter Braake, and Katja Van den Hurk. "Cardiovascular Benefits for Blood Donors? A Systematic Review". In: *Transfusion Medicine Reviews* 36.3 (July 2022), pp. 143–151.

[131] J. D. Cook and B. S. Skikne. "Iron deficiency: definition and diagnosis". In: *Journal of Internal Medicine* 226.5 (1989), pp. 349–355.

[132] D. Hugh Rushton and Julian H. Barth. "What is the evidence for gender differences in ferritin and haemoglobin?" In: *Critical Reviews in Oncology/Hematology* 73.1 (Jan. 2010), pp. 1–9.

[133] Alan E. Mast, Aniko Szabo, Mars Stone, Ritchard G. Cable, Bryan R. Spencer, Joseph E. Kiss, and for the NHLBI Recipient Epidemiology Donor Evaluation Study (REDS)-III. "The benefits of iron supplementation following blood donation vary with baseline iron status". In: *American Journal of Hematology* 95.7 (2020), pp. 784–791.

[134] Jan Karregat, Maike G. Sweegers, Franke A. Quee, Henriëtte H. Weekamp, Dorine W. Swinkels, Věra M. J. Novotny, Hans L. Zaaijer, and Katja van den Hurk. "Ferritin-guided iron supplementation in whole blood donors: optimal dosage, donor response, return and efficacy (FORTE)-a randomised controlled trial protocol". In: *BMJ open* 12.3 (Mar. 2022), e056316.

[135] Amber Meulenbeld, Steven Ramondt, Maike G. Sweegers, Franke A. Quee, Femmeke J. Prinsze, Emiel O. Hoogendijk, Dorine W. Swinkels, and Katja van den Hurk. *Effectiveness of Ferritin-guided Donation Intervals in Blood Donors: results of the Stepped Wedge Cluster-randomised FIND'EM Trial*. Jan. 2023.

[136] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. *Attention Is All You Need*. Dec. 2017.

[137] Lei Huang, Feng Mao, Kai Zhang, and Zhiheng Li. "Spatial-Temporal Convolutional Transformer Network for Multivariate Time Series Forecasting". In: *Sensors* 22.3 (Jan. 2022), p. 841.

Nederlandse samenvatting

Het onderzoek in dit proefschrift heeft als doel bloeddonatieprocessen bij Sanquin te verbeteren. Het belangrijkste gezondheidsrisico voor bloeddonors is ijzertekort, dat wordt geanalyseerd op basis van hemoglobine- en ferritineniveaus van donors. Als een van deze niveaus ontoereikend is, wordt de donor uitgesteld van donatie. Uitstel vanwege een laag hemoglobineniveau vindt ter plekke plaats, wat betekent dat de donor al naar de bloedbank is gereisd en dan zonder te doneren naar huis moet terugkeren, wat demotiverend is voor de donor en inefficiënt voor de bloedbank. Een groot deel van dit proefschrift heeft dan ook als doel een voorspellend model te ontwikkelen voor hemoglobineniveaus van donors, gebaseerd op historische metingen en donorkenmerken.

Het ontwikkelde model vermindert het uitstelpercentage met ongeveer 60% (van 3% naar 1% voor vrouwen en van 1% naar 0,4% voor mannen), wat laat zien dat het gebruik van data de efficiëntie van het beleid van bloedbanken kan verbeteren. Bovendien zijn de voorspellingen van het model verklaarbaar gemaakt, waardoor de bloedbank inzicht krijgt in waarom specifieke voorspellingen worden gedaan. Deze inzichten vergroten ons begrip van de relaties tussen donorkenmerken en hemoglobineniveaus. Als dit voorspellingsmodel in de praktijk zou worden toegepast, zouden de verklaringen ook met de donor kunnen worden gedeeld om hen te helpen begrijpen waarom ze wel of niet worden uitgenodigd om te doneren, wat ook kan bijdragen aan de tevredenheid en het behoud van donors.

In een gezamenlijke studie met bloedbanken in Australië, België, Finland en Zuid-Afrika werd hetzelfde voorspellende model toegepast op data van elke bloedbank.

Ondanks verschillen in beleid en donordemografieën leerden de modellen vergelijkbare verbanden met de voorspellende variabelen in alle landen. Verschillen in prestaties konden voornamelijk worden toegeschreven aan verschillen in uitstelpercentages, waarbij bloedbanken met hogere uitstelratio's een hogere modelnauwkeurigheid behaalden.

Naast modellen voor hemoglobinevoorspelling werden ook andere vragen onderzocht. Een studie heeft als doel determinanten van ferritineniveaus bij donors te identificeren met behulp van herhaalde metingen en koppelingen aan omgevingsvariabelen. Een andere studie betreft het modelleren van de farmacokinetiek van antilichamen tegen COVID-19 bij donors en het vinden van relaties tussen patiëntkenmerken, symptomen en antilichaamniveaus over de loop van de tijd.

Samengevat laat het onderzoek in dit proefschrift het potentieel zien binnen de rijkdom aan data die verzameld wordt door bloedbanken. De voorgestelde op data gebaseerde donatiestrategieën verminderen niet alleen het aantal uitstelgevallen, maar verhogen ook het behoud en begrip van donors. Deze aanpak stelt Sanquin in staat om meer gepersonaliseerde feedback te geven aan donors over hun ijzerstatus, waardoor het bloeddonatieproces wordt geoptimaliseerd en de algehele effectiviteit van bloedbanksystemen verbetert.

English summary

The research in this dissertation aims to optimise blood donation processes in the framework of the Dutch national blood bank Sanquin. The primary health risk for blood donors is iron deficiency, which is evaluated based on donors' hemoglobin and ferritin levels. If either of these levels are inadequate, donors are deferred from donation. Deferral due to low hemoglobin levels occurs on-site, meaning that donors have already traveled to the blood bank and then have to return home without donating, which is demotivating for the donor and inefficient for the blood bank. A large part of this dissertation therefore has the objective to develop a prediction model for donors' hemoglobin levels, based on historical measurements and donor characteristics.

The prediction model that was developed reduces the deferral rate by approximately 60% (from 3% to 1% for women, and from 1% to 0.4% for men), showing the potential of using data to enhance blood bank policy efficiency. Additionally, the model predictions were made explainable, providing the blood bank with insights into why specific predictions are made. These insights increase our understanding of the relationships between donor characteristics and hemoglobin levels. If this prediction model would be implemented in practice, the explanations could also be shared with the donor to help them understand why they are (not) invited to donate, which could also contribute to donor satisfaction and retention.

In a collaborative effort with blood banks in Australia, Belgium, Finland and South Africa, the same prediction model was applied on data from each blood bank. Despite differences in blood bank policies and donor demographics, the models found similar associations with the predictor variables in all countries. Differences in performance

could mostly be attributed to differences in deferral rates, with blood banks with higher deferral rates obtaining higher model accuracy.

Beyond hemoglobin prediction models, additional research questions are explored. One study aims to identify determinants of ferritin levels in donors through repeated measurements, and linking these to environmental variables. Another study involves modeling the pharmacokinetics of antibodies in COVID-19 recovered donors, and finding relationships between patient characteristics, symptoms, and antibody levels over time.

In summary, the research in this dissertation shows the potential within the wealth of data collected by blood banks. The proposed data-driven donation strategies not only decrease deferral rates but also increase donor retention and understanding. This comprehensive approach allows Sanquin to provide more personalised feedback to donors regarding their iron status, ultimately optimising the blood donation process and contributing to the overall efficacy of blood banking systems.

Dankwoord

Allereerst wil ik graag mijn promotieteam enorm bedanken voor al hun steun en advies in de afgelopen vier jaar (inmiddels vijf). Alledrie hebben jullie een heel andere rol gespeeld in mijn begeleiding, en hebben me verschillende lessen geleerd. Ik heb er enorm van genoten met jullie samen te werken en had me geen beter begeleidingsteam kunnen voorstellen.

Matthijs, ik voel me vereerd dat ik de eerste promovenda ben die je als officiele promotor hebt begeleid - deze rol is je absoluut op het lijf geschreven. Ik heb me altijd door jou gesteund gevoeld, en je was altijd beschikbaar als ik advies nodig had. Jouw "computer science" perspectief op mijn vaak toch erg toegepaste onderzoek was heel waardevol. Ik ben blij dat we onze samenwerking kunnen voortzetten als collega-docenten bij het LIACS.

Mart, jouw begeleiding en steun heeft me door enkele zware periodes geholpen gedurende dit promotietraject, en daar ben ik enorm dankbaar voor. Je hebt me niet alleen veel geleerd over onderzoek, maar ook in het kader van persoonlijke ontwikkeling, en hoe je alles een beetje in perspectief houdt. Mijn tijd bij Sanquin zit er helaas op, maar we verliezen elkaar zeker niet uit het oog.

Katja, ik heb erg geluk gehad met jouw betrokkenheid bij mijn onderzoek. Jouw perspectief als epidemioloog en redacteur hebben mijn onderzoek en de bijbehorende publicaties tot een hoger niveau getild. Bovendien was ik lang niet zo trots geweest op de introductie en discussie van dit proefschrift als ik jouw feedback niet had gehad.

Ik wil graag iedereen bedanken van de verschillende onderzoeksgroepen waar ik onderdeel van heb mogen uitmaken. Bedankt aan iedereen van de TTA groep bij

Sanquin: Shannon, Syeldy, Merel en Amber (als TTA erelid), het was een plezier om met jullie te werken en samen op conferentie te gaan. Bedankt ook aan iedereen van het Data Science Research Programme, in het bijzonder aan Alex, Anne, Annelieke, Daniela, Gineke, Hugo, Manon, Wout en beide Wouters. Ik heb genoten van onze murder mystery avonden en de *Among us* speelsessies tijdens de lockdowns. Een bijzonder bedankje aan Anne, die me erop wees toen het NFI een nieuwe data scientist zocht; enorm leuk dat we nu weer collega's zijn! Ik bedank ook graag mijn EDA/LIACS-collega's: Iris, Lincen, Sander en Suzan, bedankt voor alle gezellige lunches en praatjes. Thanks also to all members of the SanguinStats group for the nice international collaborations, and in particular *kiitos paljon* to Mikko and Jarkko for making my research trip to Helsinki not just productive, but also very enjoyable!

Dank ook aan alle vrienden en (schoon)familie die me gesteund en/of afgeleid hebben de afgelopen jaren. Specifiek wil ik Chava, Isabelle, Ozair en Vera bedanken voor onze online schrijfsessies tijdens de lockdown - na vandaag zijn we (hopelijk) allemaal doctor! Isa en Isabelle, enorm bedankt dat jullie mijn paranimfen willen zijn en naast me staan op deze bijzondere en mooie dag. Tot slot wil ik natuurlijk Dirk ontzettend bedanken: je was er voor me op de hoogte- en dieptepunten. Hoewel je geen woord van mijn papers hebt gelezen, heb je ik-weet-niet-hoeveel emails voor me proefgelezen en geluisterd naar al mijn succesverhalen en gefrustreerde monologen - en dat waren er nogal wat gedurende de afgelopen vijf jaar. Dankjewel voor al je steun en aanmoediging, voor je relativerende kijk op alles, en voor ons mooie leven samen.

Curriculum vitae

Marieke Vinkenoog werd geboren in Amsterdam op 16 juli 1993. Nadat zij in 2011 haar vwo-diploma behaalde aan het Vossius Gymnasium vertrok ze naar Leiden om te studeren. Ze heeft haar propedeuse Geneeskunde behaald en stapte daarna over naar de bacheloropleiding Biologie. Daarnaast volgde ze het Honours College traject, waarin ze zich enerzijds verdiepte in evolutiebiologie, anderzijds in het uitvoeren van onderzoek in het algemeen. Na de bachelor Biologie besloot ze zich verder te specialiseren in statistiek, en begon aan de masteropleiding Statistical Science for the Life and Behavioural Sciences, met de specialisatie Data Science.

In het collegejaar 2015-2016 was Marieke het studentlid van het faculteitsbestuur van de Faculteit der Wiskunde en Natuurwetenschappen. In deze rol hield ze zich onder andere bezig met de verbetering van informatievoorziening van studentleden van opleidingscommissies, en was ze mede-ontwikkelaar van een online cursus voor deze nieuwe leden. In het opvolgende collegejaar was ze voorzitter van de Stichting Bètabanenmarkt, die jaarlijks een carrière-evenement voor alle studenten van de faculteit organiseert.

Marieke behaalde haar masterdiploma cum laude in 2018, na een scriptie over het schatten van diplotypefrequenties in gestratificeerde populaties. In oktober 2018 begon zij aan haar promotietraject bij Sanquin en de Universiteit Leiden, waar dit proefschrift het resultaat van is.

Sinds januari 2023 is Marieke vier dagen per week als data scientist werkzaam bij het NFI, en één dag per week als docent bij het LIACS. Bij het NFI houdt ze zich in teamverband bezig met het ontwikkelen van datagedreven oplossingen die de forensische opsporing ondersteunen. Als docent verzorgt ze het vak Statistics for Computer Scientists voor tweedejaars informaticastudenten en de Honours Class Data Science.